TOWARD AN AUTOMATED FIRST IMPRESSION ON PATENT CLAIM VALIDITY: ALGORITHMICALLY ASSOCIATING CLAIM LANGUAGE WITH SPECIFIC RULES OF LAW

Aashish R. Karkhanis & Jenna L. Parenti*

CITE AS: 19 STAN. TECH. L. REV. 196 (2016)

EXECUTIVE SUMMARY

Can an algorithm identify words that place patent claims at higher risk of invalidation? Today's language processing software can hardly compete with a seasoned legal professional in assessing claim quality. It may, however, uncover patterns that indicate similarity of patent claim terms to others that have already been challenged in the courts. This methodology quantitatively approximates the invalidity risk a patent claim faces based on the similarity of its words to other claims adjudicated under a specific rule of law. In this way, software-based linguistic analysis according to the methodology presented here provides a starting point for efficiently and algorithmically generating a legal "first impression" on the text of a patent claim.

This study explores potential correlation of keywords to patent eligibility,[1] *a legal doctrine restricting a patent's monopoly power to innovations of particular types. This methodology explores the possibility of estimating risk for a particular claim, based on the presence of keywords commonly seen in claims previously adjudicated for validity under a specific rule of law. Such tools could efficiently reduce the scope of uncertainty for factors indicating patent quality and provide alternatives to costly litigation for patent value discovery.*[2]

* With many thanks to our colleagues at RPX Corporation for their outstanding research assistance, expertise, and enthusiasm for this study; and to Prof. Lee Fleming of the Fung Institute for Engineering Leadership at the University of California at Berkeley for his and his students' perspectives on legal analytics.

[1] 35 U.S.C. § 101 (2012) ("Whoever invents or discovers any new and useful process, machine, manufacture, or composition of matter, or any new and useful improvement thereof, may obtain a patent therefore, subject to the conditions and requirements of this title.").

[2] RPX Corp., 2014 NPE COST REPORT HIGH LEVEL FINDINGS (2015), http://www.rpxcorp.com/wp-content/uploads/sites/2/2015/05/RPX-2014-NPE-Cost-Report-ZZFinal.pdf [https://perma.cc/NU4A-3EP6] ("In the majority of NPE assertions, more than half the cost to operating companies is legal cost.").

ABSTRACT

The methodology presented here combines natural language processing software with frequency analysis to estimate patent claim validity under specific rules of law. Source data consists of a manually curated collection of patent claims adjudicated under a particular rule of law, along with manually coded symbols representing judicial outcomes for each claim under that rule of law. Natural language processing algorithms use that collection to create an index of keywords found in the patent claims' language. The index is then used to form a set of patent claim keywords with potentially heightened correlation to validity or invalidity outcomes for the rule in question. Each keyword is compared against the original collection of patent claims and outcomes, and the frequencies of the keyword's appearances in challenged claims and invalidated claims are separately aggregated. Each keyword is assigned a validity score indicating its relative prevalence in either validated or invalidated claims, and a frequency score indicating the percentage of claims in which the keyword appears. The two scores are independent, but together they indicate the correlation of a keyword to claims adjudicated as either valid or invalid under a particular rule.

An exemplary model based on this methodology explores the correlation between certain patent keywords and validity outcomes for patent eligibility under 35 U.S.C. § 101. Patent eligibility is suitable to analysis under this model because the issue, like the methodology proposed here, is focused largely on claim language rather than extrinsic factors. The model draws solely from a manually curated set of recent "post-Bilksi" patent eligibility rulings on claims challenged in U.S. district courts. Natural language processing using the Python Natural Language Toolkit ("NLTK") extracts keywords along with their contextual parts of speech for scoring. Frequency analysis of these keywords results in an automatically generated list of keywords with varying correlation to patent eligibility invalidity outcomes.

TABLE OF CONTENTS

I. BACKGROUND

Congress intended for "anything under the sun made by man"[3] to be eligible for patent protection. This broad scope of patent coverage is reflected in statute, with 35 U.S.C. § 101[4] reciting four broad categories of patent-eligible subject matter.[5] The Supreme Court has carved out three judicial exceptions[6] to these categories, holding unpatentable laws of nature, physical phenomena, and abstract ideas. In attempting to balance this wide statutory scope of patent coverage

[3] Diamond v. Chakrabarty, 447 U.S. 303, 308 (1980) ("In choosing such expansive terms . . . modified by the comprehensive 'any,' Congress plainly contemplated that the patent laws would be given wide scope.").

[4] See *supra* note 1.

[5] *In re* Bilski, 545 F.3d at 943, 951. (Fed. Cir. 2008).

[6] Chakrabarty, 447 U.S. at 309 (1980).

against its judicial exceptions, lower courts have struggled to define the line between a patentable invention and an unpatentable principle.[7] For example, the line between patentable computer technology and unpatentable abstract mathematical ideas remains blurred,[8] complicating a clear understanding of rights conveyed by certain patent claims.

As case law involving § 101 eligibility exceptions grows increasingly complex, lower courts are left without clear guidance for differentiating between eligible and ineligible patent claims. Supreme Court precedent defines notable exceptions to the statute's broad coverage. These exceptions are derived from the concern of preempting the use of natural laws,[9] and thus inhibiting further discoveries based on those laws.[10] In its recent *Alice Corp. Pty. Ltd. v. CLS Bank International* ruling, the Court cautioned against too expansive an interpretation of the § 101 eligibility exceptions, reasoning that all inventions involve underlying natural laws on some level.[11] Nonetheless, the Court ruled against the patent claims at issue in *Alice*, finding them ineligible for reciting an unpatentable abstract idea.

The patent eligibility inquiry of *Alice* follows two steps. Courts first determine whether claims are drawn to an abstract idea or mental concept,[12] and if so, second, whether the claims contain an "inventive concept" capable of transforming the abstract idea into a patent-eligible application of that idea.[13] In *Alice*,[14] the Court again avoided establishing a bright-line patent eligibility rule,[15] inviting complicated analysis around which conventionally known concepts in particular fields are unpatentable abstract ideas.[16] As a result, the tug of war

[7] Lumen View Tech. LLC v. Findthebest.com, Inc., 984 F. Supp. 2d 189, 195. (S.D.N.Y. 2013).

[8] Parker v. Flook, 437 U.S. 584, 589 (1978) ("The line between a patentable 'process' and an unpatentable 'principle' is not always clear.").

[9] Mayo Collaborative Servs. v. Prometheus Labs., Inc., 132 S. Ct. 1289, 1294 (2012).

[10] Bilski v. Kappos, 561 U.S. 593, 612 (2010).

[11] *See* Alice Corp. Pty. Ltd. v. CLS Bank Int'l, 134 S. Ct. 2347, 2354 (2014).

[12] Gottschalk v. Benson, 409 U.S. 63, 71 (1972) (holding the patent-in-suit to be invalid under § 101 because "[t]he mathematical formula involved here has no substantial practical application except in connection with a digital computer").

[13] Parker v. Flook, 437 U.S. 584, 590 (1978) ("[P]ost-solution activity, no matter how conventional or obvious in itself, can[not] transform an unpatentable principle into a patentable process.").

[14] *See Alice*, 134 S. Ct. 2347 (2014) (applying its traditional two-step patent eligibility inquiry, the Supreme Court held invalid a patent directed to an electronic technique for mitigating settlement risk as an unpatentable abstract idea).

[15] *See* Bilski v. Kappos, 561 U.S. 593, 612, 609 (2010) ("The Court once again declines to impose limitations on the Patent Act that are inconsistent with the Act's text," and, "[r]ather than adopting categorical rules that might have wide-ranging and unforeseen impacts, the Court resolves this case narrowly on the basis of this Court's decisions in *Benson*, *Flook*, and *Diehr*, which show that petitioners' claims are not patentable processes because they are attempts to patent abstract ideas.").

[16] The post-*Bilski* Court introduced the idea of "conventional steps" without defining a standard for "conventional" activity. Mayo Collaborative Servs. v. Prometheus Labs., Inc., 132 S. Ct. 1289, 1300 (2012)("[S]imply appending conventional steps, specified at a high level of

between the broad statute and the judicial exceptions continues today, leaving the courts in disagreement[17] regarding the scope and quality of patent claims.

II. PROPOSITION

Can the degree of similarity of a patent claim's text to that of adjudicated claims imply a degree of invalidation risk for the claim under a specific rule of law? A model for testing this proposition generates weighted keywords indicating similarity to adjudicated claims and leverages software-based natural language analysis. The adjudicated patent claims are manually selected from recent district court adjudications for their relevance to a particular rule of law. From this body of claims, automated natural language analysis tools extract individual words, as well as the frequencies at which those words correlate to particular outcomes under the particular rule of law. Each word is assigned two independent numerical scores derived from these extracted values, representing the word's tendency to appear in invalid claims, and its tendency to appear throughout the selected set of claims regardless of outcome, respectively.

III. METHODOLOGY

This model creates a score-ranked list of patent claim keywords relevant to a particular rule of law. Using software-based natural language processing ("NLP"), each keyword is associated with a validity score and a frequency score. The validity score represents correlation between that keyword and judicial opinions adverse to claims containing the keyword. The frequency score independently represents the keyword's prevalence in all adjudicated claims without regard to outcome. Keywords with higher magnitude negative validity scores indicate higher correlation to judgments of patent invalidity under a specific rule of law. Though this methodology focuses on individual keywords for simplicity, more robust models based on this methodology may implement a more complicated phrase-level contextual analysis supported by NLP software.[18]

Application of the model to a target claim requires identifying which words in the claim are present in the keyword list, and aggregating keyword scores to

generality," to a method already well known in the art, is not enough to supply the inventive concept).

[17] *See* Ultramercial, Inc. v. Hulu, LLC, 772 F.3d 709, 711, 717 (Fed. Cir. 2014) (holding that advertising patent was invalid under § 101 after holding twice before that the claims were patent-eligible, even though the computer-based methods were not well-known economic or commercial practices, as in *Alice*).

[18] *See* NLTK Project, *nltk Package*, NLTK 3.0 DOCUMENTATION (Oct. 26, 2015), http://www.nltk.org/api/nltk.html [perma.cc/7YU7-ED6H] ("Finding collocations requires first calculating the frequencies of words and their appearance in the context of other words. Often the collection of words will then requiring filtering to only retain useful content terms. Each ngram of words may then be scored according to some association measure, in order to determine the relative likelihood of each ngram being a collocation.").

arrive at a claim score. Each keyword's validity score implies a degree of similarity to previously adjudicated claims. As the number of keywords with high magnitude negative validity scores increases, the claim's similarity to invalid claims increases. This similarity suggests increased invalidity risk for the target claim.

A. *Identifying Key Claims from Judicial Opinions Addressing a Specific Rule of Law*

This model supposes a set of patent claims drawn from judicial opinions addressing claim validity under a particular rule. Text analysis is limited to patent claims relevant to a particular legal issue. Limiting the set of patent claims for analysis to only those adjudicated under a particular rule may reveal patterns specific to distinct questions of law. The selection of a coherent and reasonably large set of patent claims relevant to a particular rule is a critical first step to uncovering these patterns. Under the present model, analysis is limited to rules of law depending largely on claim language. This attempts to control the effect of extrinsic factors (e.g., prior art) that could cause significant distortion in claim similarity analysis if ignored.

Human selection of judicial opinions feeds an automated linguistic analysis of the claims associated with those opinions. The model discussed here selects only judicial opinions based on current legal reasoning, to minimize the effect of patent claim language adjudicated under inaccurate or overturned reasoning. Patent claim text associated with a rule of law is drawn from patents referenced in a manually curated set of judicial opinions germane to that rule. If a judicial opinion in the set includes a ruling regarding a particular patent, the text of that patent is included in a collection of patent claim text. In this manner, the collection includes text of all claims associated with that rule of law.

B. *Identifying Keywords by Algorithmically Processing Key Claim Language*

Limiting text processing to patent claims associated with a single rule increases the possibility of uncovering patterns between claim language and a discrete rule of law. While manual text analysis at large scale is often impractical, automatic text analysis using NLP software can efficiently identify unique words in particular parts of speech from a collection of patent claims. These unique keywords, extracted from patent claims in view of grammatical context, may imply a correlation to particular rules of law.

NLP tools, including certain NLP software packages, can efficiently extract keywords and identify their associated linguistic characteristics from the corpus of patent claims.[19] Analytics tools leveraging these new software systems can classify

[19] Natural Language Toolkit ("NLTK"), provided in the Python programming language, is one of a number of software tools designed to efficiently parse and manipulate human

aspects of words at a substantially higher speed than manual reading. Such tools approximate and simulate, rather than replicate, human understanding of text.

With the ability to understand parts of speech, subtle but important variations in linguistics may be accounted for and analyzed with a high degree of specificity. For example, the verb "live," meaning "to be alive," may be algorithmically classified as distinct from the adjective "live," meaning "having life." Preserving relationships between words and parts of speech preserves some grammatical aspects of analyzed text. Given the legally operative nature of patent claim language, this additional degree of detail may offer more insight into the motivation behind the selection of certain words in patent claims than grammar-agnostic text searches.

A set of keywords from the corpus of relevant patent claims may include a single instance of each unique word present in the corpus, or some linguistically relevant subset. Particular implementations might disregard words in speech with less relative potential legal effect (e.g., articles), or might emphasize words in parts of speech with higher potential legal effect (e.g., nouns, verbs). Plausible implementations might incorporate one, both, or none of these filters, or other desired weighting or scaling of words depending on their grammatical state. With a natural language processing system capable of preserving grammatical information for a keyword, such variations may be set manually or adaptively as desired.

C. *Quantifying Relevance of Each Keyword to the Relevant Law*

The relationship between claim text and the resulting scores may be represented numerically. This methodology defines the score of each keyword with two independent values represented as a pair: a validity score, representing the keyword's tendency to indicate validity or invalidity, and a frequency score, representing the frequency with which the word appears in the set of adjudicated claims, comprise the pair of values. Thus, a keyword score contains two independent values indicating both how likely the word is to appear in an invalidated claim and how often the word is present in a challenged claim. This methodology counts a word once if it appears anywhere in a claim. Words are not counted more than once if they appear multiple times in a claim. Disregarding the number of occurrences within a claim attempts to minimize distortion based on common words in the patent lexicon (e.g., "wherein").[20]

language text. *See* NLTK Project, *Natural Language Toolkit*, NLTK 3.0 DOCUMENTATION (Oct. 26, 2015), http://www.nltk.org [https://perma.cc/28XT-C6YS].

[20] For example, this methodology counts the presence of the term "wherein" only once in claim 1 of U.S. Patent No. 4,923,450, though it appears two times in the claim. "1. A medical tube for placement into a patient comprising metal containing powdered zeolite, *wherein* at least one of the metals contained in said zeolite is substituted by at least one ion exchangeable metal selected from the group consisting of Ag, Cu and Zn, having anti-bacterial properties, the

A keyword score (“*K*”) is composed of two independent parts, representing the keyword’s relative invalidity (“Δk_w”) and frequency (“f_w”) levels with respect to the set of claims. Each keyword score is calculated from the number of occurrences of a word in various contexts, normalized for the size of the set. Unique among factors in this model, the total number of relevant rulings (“N_{total}”) remains constant for all keywords (“w”).[21]

A keyword’s (“*w*”) validity score (“Δk_w”) is the difference between the number of times the keyword appears in an invalidated claim (“$n_{invalid,w}$”) and the number of times the keyword appears in a valid claim (“$n_{valid,w}$”), divided by the number of times the keyword appears at least once in any claim (“$n_{all,w}$”). Resulting validity scores range between 1 and -1 and indicate varying levels of correlation to validity or invalidity outcomes. Positive values indicate a positive correlation with validity rulings; the higher a keyword’s validity score, the more often that keyword appeared in a claim held patent eligible. A validity score of 1 indicates perfect correlation to patent claim validity; every adjudicated claim in the set including the word was found eligible. Similarly, a validity score of -1 indicates the opposite: every claim including that keyword was found invalid. In the middle, keywords with validity scores of 0 are not correlated to validity or invalidity outcomes; they appear exactly as frequently in claims held valid as they do in claims held invalid.

$$\Delta k_w = \frac{n_{valid,w} - n_{invalid,w}}{n_{all,w}}$$

Equation 1: Keyword Score, Validity

A keyword’s frequency score is the number of rulings in which the keyword appeared at least once in a claim (“$n_{all,w}$”), divided by the total number of relevant rulings (“N_{total}”). Resulting frequency scores range between 1 and 0 and indicate varying levels of saturation into the claim set (“*C*”). A frequency score of 1 indicates complete saturation in the claim set; i.e., the keyword appears at least once in every claim in the set. Similarly, a frequency score of 0 indicates that the keyword appears in none of the examined claims in the set.

$$f_w = \frac{n_{all,w}}{N_{total}}$$

Equation 2: Keyword Score, Frequency

Each keyword’s relationship to valid and invalid claims is represented two-dimensionally, by incorporating the validity score and the frequency score into a pair. By independently representing invalidity and frequency, this model attempts to minimize distortion by claim terms that appear frequently in the set of curated claims, but also have little effect legally or even linguistically. The frequency score

zeolite being coated onto or kneaded into said medical tube *wherein* said medical tube, when in contact with said patient, continuously exhibits said antibacterial properties of said ion exchangeable metal for a period of time of at least two days.” (emphasis added).

[21] *See supra* Part III, Subpart A.

thus provides a path for objectively filtering out "noisy" keywords common in the set of patent claims but appearing roughly as often in both valid and invalid claims, without requiring subjective removal of keywords that may bias the model. The pair of scores constructed in this way describes a keyword's relationship to validity outcomes, together with its relationship to the data set as a whole.

$$K(w) = (\Delta k_w, f_w) = \left(\frac{n_{valid,w} - n_{invalid,w}}{n_{all,w}}, \frac{n_{all,w}}{N_{total}} \right)$$

Equation 3: Keyword Score, Combined

D. *Observing Similarity of a Target Claim's Words to That of Identified Keywords*

A set of keywords with associated scores can be a tool to measure the similarity of a target claim's language to terms with known associations to adverse judicial outcomes, on a word-by-word basis. All unique words appearing in a claim or collection of claims compose a set of scored keywords. A set of words in a target claim is composed of all of the words that appear in that claim. A risk score representing a claim's similarity to previously adjudicated claims is an aggregation of all scores of keywords in that claim. This way, the aggregated scores of individual words can indicate the score of a claim as a whole.

This study does not address optimal solutions or a methodology for aggregating keyword scores into risk scores. Validity score summation is a plausible (and simple) technique for generating a risk score for a target claim. Each keyword may be compared against words in the target claim. If the keyword appears at least once in the target claim, the keyword's score is aggregated into the claim's risk score. Thus, a target claim's risk score reflects both a level of similarity to other claims (based on the number of shared keywords) and a level of relevance to adverse judicial outcomes (based on the aggregated weights for each keyword). Aggregation techniques may, of course, vary depending on the particular model implemented.

IV. EXAMPLE MODEL

The model discussed here identifies keywords that are highly correlated with particular patent eligibility rulings at the district court level. Arguments against patent eligibility, in the aggregate, have recently been effective[22] in invalidating patents. Since the legal inquiry into patent eligibility focuses on the reading of the patent claims, with little dependence on extrinsic evidence, the methodology proposed here similarly focuses on a contextual analysis of patent claim text. The

[22] *See* Aashish R. Karkhanis, *Quantifying Patent Eligibility Judgments*, 15 WAKE FOREST J. BUS. & INTELL. PROP. L. 203 (2014). Recent patent eligibility challenges under both Supreme Court and Federal Circuit theories have resulted in invalidity rulings due to ineligibility under 35 U.S.C. § 101 with over a fifty percent success rate.

model quantifies how often unique claim terms appear in patent eligibility rulings, and how often those unique claim terms appear in claims held eligible and ineligible for patent protection. The presence in a target claim of keywords scored this way indicate the target claim's similarity to other patent claims that are vulnerable to validity challenge under the same rule.

A. *Selecting a Rule of Law Suited to this Methodology*

Because this methodology focuses solely on analysis of claim text language, its applications should be limited to legal rules that also rely heavily on claim text analysis. This example model explores claim text as it relates to patent eligibility. Other patent-related legal inquiries, such as weighing evidence of prior invention[23] (e.g., "prior art") or the full permissible scope of particular claim terms[24] would be less suited to analysis under this methodology because they depend on extrinsic evidence.

B. *Selecting a Corpus of Patent Claims Adjudicated Under Current Jurisprudence*

To generate linguistic correlations reflecting the highest agreement with current judicial consensus, this model considers a limited subset of all patent eligibility rulings: judgments made after the Supreme Court's decision in *Bilski v. Kappos.*[25] *Bilski* marks the beginning of the current era in patent eligibility reasoning. From a procedural perspective, the model considers rulings only at the summary judgment stage. Patent eligibility is traditionally, but not exclusively,[26] addressed at summary judgment. Though this model limits time periods and procedural stages to maximize stability,[27] other models may vary these factors under this methodology.

The NLTK-based Python program takes as input a manually generated table containing the text of all relevant claims and their associated judicial outcomes. An exemplary input data source comprises a two column by unlimited row ("2-by-N") table. At minimum, the input table according to this methodology includes a first column value ("r_t") for the text of an adjudicated claim in a particular ruling, and a second column value ("r_E") indicating whether the district court held

[23] *See* 35 U.S.C. §§ 102, 103.

[24] *See* Markman v. Westview Instruments, Inc., 517 U.S. 370 (1996).

[25] *See* Bilski v. Kappos, 561 U.S. 593, 612 (2010) (holding unpatentable as directed to an abstract idea a claim directed to automated trading of a commodity based on market risk factors).

[26] *See, e.g.*, Tuxis Tech., LLC v. Amazon, Inc., No. 13-1771-RGA, 2014 WL 4382446 (D. Del. Sept. 3, 2014). Though most judgments on patent eligibility are at summary judgment, courts have increasingly adjudicated this issue at the complaint stage. Judgment for patent eligibility on the pleadings is a recent phenomenon. Rulings at this stage are excluded from this exemplary model, in order to minimize distortion from unstable judicial procedure.

[27] *See infra* Part VII.

that claim eligible or ineligible under 35 U.S.C. § 101.[28] The judicial ruling is represented by a binary value, 1 for a ruling that the claim is patent eligible and 0 for a ruling that the claim is patent ineligible.

$$C = \{(r_t, r_E) | r_E \in \{0,1\}\}$$

Equation 4: Claim Table Set

For each relevant judicial outcome, patent claim text and binary eligibility outcomes are tracked in the first and second columns, respectively. Each row in the table corresponds to a separately argued opinion regarding a particular patent claim's eligibility. Thus, if a judicial opinion adjudicates multiple claims using a single line of reasoning for purposes of determining patent eligibility, the model represents that outcome as a single row. That row includes the text of the claim on which the court focused its analysis, and whether the court held that claim eligible or ineligible. From this table of issue-specific patent claim text and the binary-coded issue outcomes, natural language processing under this model can extract linguistic patterns tied specifically to patent eligibility.

C. *Identifying Unique Keywords in the Corpus with Python Natural Language Processing*

An analytic tool capable of efficiently parsing patent claim text for context adds insight beyond traditional text matching analyses. The Python programming language[29] supports robust natural language processing using the Python Natural Language Toolkit ("NLTK").[30] The example model discussed here leverages NLTK functionality,[31] allowing automatic association of a word with a tag[32] representing its contextual part of speech. This allows a more nuanced approximation of that word's potential legal effect.[33]

This model uses Python NLTK to create a list of unique keywords from all relevant patent claims. First, all claims are combined into a single text, represented as a Python corpus object.[34] Second, part of speech tagging functions

[28] *See supra* Part IV, Subpart A.

[29] *See* Python Software Found., *Welcome to Python*, https://www.python.org/ [https://perma.cc/4DC4-538V].

[30] NLTK Project, *Natural Language Toolkit*, NLTK 3.0 DOCUMENTATION (Oct. 26, 2015), http://www.nltk.org [https://perma.cc/28XT-C6YS] ("NLTK is a leading platform for building Python programs to work with human language data.").

[31] Python uses Penn Treebank as its default NLP engine. *See* NLTK Project, *nltk.tag package*, NLTK 3.0 DOCUMENTATION (Oct. 26, 2015), http://www.nltk.org/api/nltk.tag.html [https://perma.cc/GF27-4ATW] ("An off-the-shelf tagger is available. It uses the Penn Treebank tagset.").

[32] *See* Beatrice Santorini, *Part-of-Speech Tagging Guidelines for the Penn Treebank Project (3rd Revision)*, TECHNICAL REPORTS (CIS), July 1, 1990, at 6, http://repository.upenn.edu/cis_reports/570 [https://perma.cc/HGX2-9UCF].

[33] *See supra* Part III, Subpart B.

[34] STEVEN BIRD, EWAN KLEIN & EDWARD LOPER, NATURAL LANGUAGE PROCESSING WITH

in the NLTK[35] apply a contextual part of speech tag to each word in the single text. This permits separate analysis of words with identical spellings but different contextual uses. This tagging process tokenizes[36] each word into a pair including the word and its associated part of speech.

As in this model, contextualized words may be filtered before scoring to optimize a relative distribution of potentially legally relevant words. Filtering by removing certain contextualized words may reduce noise from common words lacking substantial legal meaning (e.g., articles). Models with strict filtering may remove words in all but the most substantial parts of speech (e.g., nouns, verbs, adjectives), while those with more permissive filtering may exclude only words in certain low interest parts of speech, if any.[37]

The model discussed here takes a more inclusive approach to present a larger universe of potentially interesting words while attempting to remove words with the highest potential to distort patent claim scoring. Here, every word with a non-alphabetical tag, and any word tagged as an article,[38] preposition, or subordinating conjunction is removed from the keyword set. Keywords remaining in the set are scored for patent eligibility.

D. *Scoring Each Keyword for Validity by Correlation to Specific Patent Eligibility Outcomes*

This model independently calculates a validity score and a frequency score for each keyword drawn from the claim table's data. From the input claim table, the

PYTHON ch. 2 (2014), http://www.nltk.org/book/ch02.html [https://perma.cc/R6MT-T3MG] ("Practical work in Natural Language Processing typically uses large bodies of linguistic data, or corpora.") (emphasis removed).

[35] STEVEN BIRD, EWAN KLEIN & EDWARD LOPER, NATURAL LANGUAGE PROCESSING WITH PYTHON ch. 5 (2014), http://www.nltk.org/book/ch05.html [https://perma.cc/48C9-R243] (The process of classifying words into their parts of speech and labeling them accordingly is known as part-of-speech tagging, POS-tagging, or simply tagging. Parts of speech are also known as word classes or lexical categories. The collection of tags used for a particular task is known as a tagset.") (emphasis removed).

[36] STEVEN BIRD, EWAN KLEIN & EDWARD LOPER, NATURAL LANGUAGE PROCESSING WITH PYTHON ch. 3 (2014), http://www.nltk.org/book/ch03.html [https://perma.cc/EN65-4E82] ("[W]e want to break up the string into words and punctuation . . . [t]his step is called tokenization, and it produces our familiar structure, a list of words and punctuation.") (emphasis removed).

[37] Though NLTK includes a group of common words with little lexical meaning in general ("stopwords"), this example model does not leverage that library in order to err on removing words conservatively. *See* STEVEN BIRD, EWAN KLEIN & EDWARD LOPER, NATURAL LANGUAGE PROCESSING WITH PYTHON ch. 2 (2014), http://www.nltk.org/book/ch02.html [https://perma.cc/R6MT-T3MG].

[38] The Penn Treebank schema identifies each article (e.g., "the") in text as a "determiner" with the tag code "[DT]." *See* Beatrice Santorini, *Part-of-Speech Tagging Guidelines for the Penn Treebank Project (3rd Revision)*, TECHNICAL REPORTS (CIS), July 1, 1990, at 3, 6, http://repository.upenn.edu/cis_reports/570 [https://perma.cc/HGX2-9UCF].

model generates an output keyword table[39] including two keyword columns, three raw data columns, and two calculated data columns for each keyword. The keyword columns respectively hold the keyword itself and its contextual part of speech,[40] while the five data columns quantify the keyword's score ("*K*").[41]

The three raw data columns hold numbers of occurrences and validity for each keyword. The three columns contain the number of invalidated claims in which the keyword appears ("$n_{invalid,w}$"), the number of claims held valid in which the keyword appears ("$n_{valid,w}$"), and the number of claims in which the keyword appears at least once ("$n_{all,w}$"), respectively.[42] The first calculated data column holds the keyword's validity score ("Δk_w"), and the second its frequency score ("f_w"). Each keyword's validity and frequency scores are derived from the values in its three raw data columns and the total number of rulings ("N_{total}"),[43] which corresponds to the number of rows in the claim table ("$|C|$") in this example model.

V. EXPERIMENTAL RESULTS

The algorithmically generated set of 1,122 keywords spans a wide range of keyword scores, frequency scores, and parts of speech. Although many keywords have high validity scores by absolute value, and many others have high frequency scores, a relatively small subset has high scores in both categories. To create a one-dimensional ranking from a two-dimensional list, this model ranks keywords in order first by validity score, and second by frequency score. This ranking system emphasizes words with particular outcomes for patent eligibility under the assumption that the number of occurrences is less insightful due to the small sample size of the set of claims. All keywords with the greatest negative validity scores are ranked more highly than those with more positive validity scores, while keywords having equal validity scores are sorted by their frequency scores. As a result, the model presented here weighs most heavily a keyword's relative tendency to appear in either invalid or valid claims.

[39] *See infra* Appendix B.

[40] *See supra* Part IV, Subpart C.

[41] *See supra* Equation 3.

[42] *See supra* Part III, Subpart C.

[43] *See supra* Part III, Subpart C.

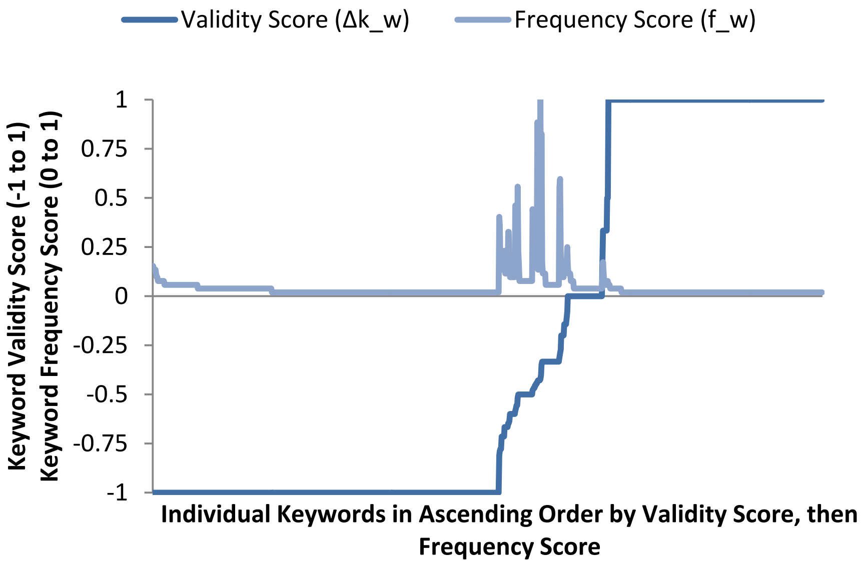

Figure 1: Automatically Generated Keywords by Score

A. *A Large Percentage of Keywords Have High Validity Scores*

Validity scores for the set of keywords are clustered at the positive and negative ends of the range. Of the 1,122 scored keywords algorithmically generated by this example model, 695 keywords have validity scores less than zero, indicating a tendency to appear in patent ineligible claims, 369 have validity scores greater than zero, indicating a tendency to appear in patent eligible claims, and 58 have validity scores equal to zero, indicating a tendency to appear just as often in patent eligible claims as patent ineligible claims. These results imply that more keywords exist that tend to appear in patent ineligible claims than eligible, possibly illustrating self-selection of claims vulnerable to patent eligibility challenge all the way to adjudication.

More telling than the general distribution above and below zero is the substantial percentage of keywords with validity scores of either 1 or -1. Of the 1,122 scored keywords, 580 have validity score equal to -1, indicating a presence only in claims ruled patent ineligible, and 359 have scores equal to 1, indicating a presence only in claims ruled patent eligible. This model presumes that keywords with validity scores at either extreme should strongly correlate to either affirmative or adverse patent eligibility outcomes.

B. *A Small Percentage of Keywords Have High Frequency Scores*

Frequency scores for the set of keywords are clustered in the bottom of the range. Of 1,122 scored keywords algorithmically generated by this example model, only one keyword (the conjunction "and") appeared in all adjudicated claims. At lower frequency score thresholds, six keywords appear in over 50% of claims, fifteen keywords appear in over 33% of claims, and twenty-six keywords appear in over 20% of claims. The bulk of keywords appear in few claims in the set. Almost two-thirds of keywords, numbering 718, appear in only one claim. Over four-fifths of keywords, numbering 909, appear in a maximum of two claims. Most keywords are thus present in only a small subset of the total claims.

C. *A Small Percentage of Keywords Have High Validity and High Frequency Scores*

In this example model, the top 100 words are presumed to have the highest degree of correlation to adverse patent eligibility outcomes as compared with the remainder of the keyword set. The model orders the list of keywords first by validity score in ascending order, and second by frequency score in descending order. A keyword's validity score is emphasized over the keyword's frequency score in order to emphasize words with disproportionate presence in claims held patent ineligible, and to deemphasize keywords more evenly distributed between valid and invalid claims throughout the set.

The model presented here generally assumes that a validity score is a much stronger indication of similarity to previously invalidated claims than a frequency score. As a result, keywords with low magnitude frequency scores but maximum validity scores of either -1 or 1 are ranked more highly than keywords that may have slightly lower validity scores and higher frequency scores. Other models based on this methodology may weight validity scores and frequency scores differently to arrive at a combined score indicating a keyword's correlation to invalidity outcomes.

A number of notable claim terms with the highest frequency scores with respect to patent eligibility are directed to particular technology areas. For example, software terminology (e.g., "processor" and "digital") and business method terminology (e.g., "transactions" and "price") appear in the top-scoring keywords. This automatic detection appears consistent with recent Supreme Court rulings on patent eligibility in these technical fields.[44]

Further information can be gleaned from studying the top twenty keywords ranked by this methodology. Grammatically, nouns compose twelve of the top twenty keywords, with verbs composing five of the top twenty, and adjectives composing two. This keyword list also includes keywords counted distinctly as

[44] *See* Alice Corp. v. CLS Bank Int'l, 134 S. Ct. 2347, 2360 (2014) (holding invalid as patent ineligible a claim generally directed to software); Mayo Collaborative Servs. v. Prometheus Labs., 132 S. Ct. 1289, 1294, 1305 (2012) (holding invalid as patent ineligible a claim generally directed to pharmaceuticals); Bilski v. Kappos, 561 U.S. 593, 612 (2010) (holding invalid as patent ineligible a financial accounting technique).

having different parts of speech.

Keyword (w)	POS	$n_{all,w}$	$n_{invalid,w}$	$n_{valid,w}$	Validity Score (Δk_w)	Frequency Score (f_w)
output	NN	8	8	0	-1	0.154
part	NN	7	7	0	-1	0.135
price	NN	7	7	0	-1	0.135
processor	NN	7	7	0	-1	0.135
transactions	NNS	7	7	0	-1	0.135
input	NN	6	6	0	-1	0.115
group	NN	5	5	0	-1	0.096
property	NN	5	5	0	-1	0.096
describing	VBG	4	4	0	-1	0.077
digital	JJ	4	4	0	-1	0.077
identifying	VBG	4	4	0	-1	0.077
independent	NN	4	4	0	-1	0.077
less	JJR	4	4	0	-1	0.077
performed	VBN	4	4	0	-1	0.077
rate	NN	4	4	0	-1	0.077
record	NN	4	4	0	-1	0.077
retrieval	NN	4	4	0	-1	0.077
sending	VBG	4	4	0	-1	0.077
unit	NN	4	4	0	-1	0.077
allowing	VBG	3	3	0	-1	0.058

Figure 2: Top 20 Keywords for Patent Eligibility, Sorted by Validity and Frequency Scores

VI. FEATURES AND FLEXIBILITY

This methodology and associated example model are intended to provide general directional guidance regarding the similarity of a given claim to previously adjudicated claims under a particular legal rule. By substantially simplifying a nuanced legal inquiry into a rules and grammar based text analysis, the model seeks to enable the creation of broad tiers of potential risk for a particular claim. Additionally, the methodology proposed here can flexibly adjust to changing consensus opinions regarding the validity of patent claims. Thus, the methodology combines broad guidance with up-to-date inputs to tier patent claims according to invalidity risk.

A. *Methodology Simulates a Legal "First Impression"*

Automated claim analysis based on the methodology proposed here creates an approximation of claim invalidity risk based on the similarity of a claim's language to claims that have been challenged, successfully or unsuccessfully, under a given rule of law. The similarity proposed here does not imply a complete duplication or replacement of human legal analysis of a particular claim. The model does, however, attempt to provide directional guidance as to whether a patent claim may be at substantial or insubstantial invalidation risk based on its risk score.[45] Quickly categorizing a general risk level may be useful for analysis of claim language where a full adjudication is not needed to assess or understand value.

This technique is focused on, and limited to, analysis of patent claims under rules of law that rely extensively on patent claim text. Certain inquiries into patent claim quality rely largely on patent claim text,[46] while others rely much more substantially on extrinsic evidence.[47] This methodology attempts to simplify a first-pass analysis for legal inquiries with minimal reliance on extrinsic evidence. When textual analysis alone is insufficient to determine claim quality, the model attempts to avoid overreaching into proposing outcomes for claims.

B. *Model Evolves with Changing Jurisprudence*

The model suggested here is designed to remain up-to-date with current jurisprudence. Users may manually input as many or as few relevant claims as desired, irrespective of adjudication date. The exemplary model's flexibility allows analysis of an open-ended corpus of patent claims drawn manually from claim adjudications under a particular legal rule.

VII. LIMITATIONS AND CONCERNS

This model relies on the assumption that detecting the presence of a select group of keywords may estimate invalidation risk under § 101 for a particular patent claim. It also assumes that claims containing similar terms will generate similar validity results in district courts. Under this assumption, an algorithmically generated list of keywords whose validity scores are generated

[45] *See supra* Part III, Subpart C.

[46] *See* 35 U.S.C. § 101 (2012).

[47] *See* 35 U.S.C. §§ 102, 103.

using litigation data from recent cases forms the basis for the model's predictive analysis. The methodology introduces potential bias due to the relatively limited body of legal data that forms the basis for the model. It may also be limited by instability in the jurisprudence caused by disagreement among the courts regarding patent eligibility issues.

A. *Limited Available Data May Affect Reliability of Observations*

Because patent eligibility rulings are relatively infrequent, the legal data available to power the model is limited. Under the described methodology, the data set available is further narrowed to the subset of patent eligibility rulings in the post-*Bilski* era that reached the summary judgment stage of a district court case. The use of this narrower data set could potentially affect the model's ability to make reliable estimations. The focus on current cases to minimize distortion caused by cases adjudicated under dated law may reduce the model's dependability. To maximize stability, this model restricts data to district court adjudications during a particular time period and a particular stage of adjudication. Other models employing this methodology may analyze as input a more diverse legal corpus.

By focusing on the individual words in each relevant claim, however, the size of the data set increases. Though 1,122 scored keywords[48] in this example model may not constitute a large data set in absolute terms, that set is a notable twenty-fold increase over the original 53 judicial opinions identified for analysis.[49] This keyword-specific analysis thus increases available data through increased granularity, as compared to analysis limited to binary validity outcomes.

B. *Instability in Legal Data from Conflicting Judicial Views May Limit Predictive Power*

Litigation data is sourced from unstable current legal guidance on patent eligibility. The courts are in disagreement regarding patent eligibility rules and they struggle to balance upholding the broad language of the statute with respecting the judicially created exceptions.[50] This instability may lead to new, potentially overruling, precedent. Such changes in law could render previous judicial outcomes bad law, and limit the number of existing adjudications fit for ingestions intoa model based on this methodology. A substantial reduction in available case law may negatively impact the predictive ability of the model and lessen its use as a tool to reduce uncertainty around patent eligibility.

[48] *See supra* Part V.

[49] *See infra* Appendix A.

[50] Diamond v. Chakrabarty, 447 U.S. 303, 309 (1980).

VIII. CONCLUSION

Estimating patent eligibility risk quantitatively for individual claims would allow human experts to efficiently sort patent claims into risk tiers in preparation for more sophisticated human analysis. The model discussed here, which focuses solely on claim text, uses a generated set of patent claim keywords ranked using a numerical validity score representing correlation to prior ineligibility outcomes, as well as a numerical frequency score representing prevalence in the set of adjudicated claims examined. As the presence of keywords with high invalidity and frequency scores in a particular target claim increases, that target claim's degree of similarity to known patent ineligible claims also increases. Thus, the methodology and example model presented here examine the correlation between terms appearing in a target patent claim and existing patent-eligibility outcomes for claims with similar claim terms. Approximating claim similarity might help clarify the scope of a patent claim's coverage and potentially reduce unproductive patent litigation.

IX. APPENDIX A: RELEVANT PATENT CLAIMS AND ASSOCIATED RULINGS

All litigations analyzed in this study appear below. Input claim text data is drawn from the full text of patent claims adjudicated for patent eligibility at summary judgment. The row for each relevant ruling includes the case name, the docket entry containing the court's reasoning at summary judgment, the ruling on eligibility, the patent at issue, and the representative disputed claim. Where multiple entries exist for a single case, the court has provided more than one independent line of reasoning for or against the eligibility of particular patent claims.

Case Name	SJ Dkt No.	Eligible?	Patent No.	Claim No.
Applied Innovation Inc v. Commercial Recovery Corp.	103	Y	7,167,839	1
AutoForm Engineering GmbH v. Engineering Technology Associates, Inc.	109	Y	7,894,929	1
Bascom Research, LLC v. Facebook, Inc.	131	N	7,139,974	45
CMG Financial Services Inc v. Pacific Trust Bank FSB	164	N	7,627,509	1
Comcast IP Holdings I LLC v. Sprint Communications Company LP	291	N	6,873,694	21
Digitech Image Technologies LLC v. Pentax Ricoh Imaging Co. Ltd.	36	N	6,128,415	1
Digitech Image Technologies LLC v. Pentax Ricoh Imaging Co Ltd.	36	N	6,128,415	1
East Coast Sheet Metal Fabricating Corp. v. Autodesk, Inc.	193	N	7,917,340	1
Enfish LLC v. Microsoft Corp.	303	N	6,151,604	47
Enfish LLC v. Microsoft Corp.	303	N	6,151,604	17
Hemopet v. Hills Pet Nutrition Inc	119	N	7,865,343	1
McRO Inc. v. BANDAI NAMCO Games America Inc	365	N	6,307,576	1
Open Text S.A. v. Box, Inc.	454	N	6,223,177	1
Steve Morsa v. Facebook Inc.	66	N	7,904,337	12
MyMedicalRecords v. Walgreen	104	N	8,301,466	8
Celsis In Vitro, Inc. v. Cellzdirect, Inc.	429	N	7,604,929	1
Genetic Veterinary Sciences, Inc. v. Canine Eic Genetics, LLC	58	N	8,178,297	1
DDR Holdings v. Hotels.com	601	Y	7,818,399	19
Accenture Global Services v. Guidewire Software, Inc.	527	N	7,013,284	1
Advanced Software Design v. Federal Reserve Bank St Louis	296	Y	6,792,110	1
Ariosa Diagnostics, Inc. v. Sequenom, Inc.	254	N	6,258,540	1
Bancorp Services v. Sun Life Assurance	396	N	5,926,792	9
Big Baboon Corp. v. Dell	384	Y	6,115,690	36
Chamberlain Group, Inc. v. Lear Corp.	810	Y	6,154,544	1

Case Name (Continued)	SJ Dkt No.	Eligible?	Patent No.	Claim No.
CLS Bank International v. Alice Corp.	104	N	7,149,720	1
CLS Bank International v. Alice Corp.	104	N	5,970,479	33
CyberFone Sys. LLC v. CNN Interactive Grp. Inc.	195	N	8,019,060	1
DietGoal Innovations LLC v. Bravo Media LLC	148	N	6,585,516	12
DietGoal Innovations LLC v. Bravo Media LLC	148	N	6,585,516	1
Digitech Image Tech. LLC v. Electronics For Imaging Inc.	88	N	6,128,415	1
Every Penny Counts, Inc. v. Wells Fargo Bank, N.A.	68	N	7,571,849	1
Federal Home Loan Mortgage Corp. v. Graff/Ross Holdings	25	N	7,908,202	104
Federal Home Loan Mortgage Corp. v. Graff/Ross Holdings, LLP	44	N	7,908,202	3
Federal Home Loan Mortgage Corp. v. Graff/Ross Holdings, LLP	44	N	7,908,202	98
Federal Home Loan Mortgage Corp. v. Graff/Ross Holdings, LLP	44	N	7,685,053	1
Federal Home Loan Mortgage Corp. v. Graff/Ross Holdings, LLP	44	N	7,685,053	51
France Telecom S.A. v. Marvell Semiconductor, Inc.	159	Y	5,446,747	1
Fuzzysharp Tech. Inc. v. Intel Corp.	76	N	6,618,047	1
Intellectual Ventures I LLC v. Capital One Fin. Corp.	371	N	8,083,137	5
Island Intellectual Prop., LLC v. Promontory Interfinancial	265	Y	7,519,551	18
LML Patent Corp. v. JP Morgan	650	Y	RE40220	67
CyberFone Sys. LLC v. ZTE (USA) Inc.	274	N	8,019,060	1
CyberFone Sys. LLC v. American Airlines Inc.	167	N	8,019,060	1
Nazomi Communication Inc. v. Samsung Telecommnications	161	Y	6,338,160	1
Oleksy v. General Electric Co.	382	Y	6,449,529	1
Oplus Tech., Ltd. v. Sears Holdings Corp.	113	Y	6,239,842	1
PerkinElmer, Inc v. Intema Ltd.	277	Y	6,573,103	1
Planet Bingo, LLC v. VKGS, LLC	73	N	6,398,646	1
Planet Bingo, LLC v. VKGS, LLC	73	N	6,398,646	1
Prompt Medical Systems LP v. AllscriptsMysis Healthcare Solutions, Inc.	410	Y	5,483,443	1
Smartgene, Inc. v. Advanced Bio. Laboratories, SA	66	N	6,081,786	1
TQP Development, LLC v. Intuit Inc.	150	Y	5,412,730	1

X. APPENDIX B: AUTOMATED GENERATED KEYWORDS (TOP 100)

The 100 keywords with the highest frequency scores appear below. Keywords are ranked in descending order first by validity score, and second by frequency score. Each row includes an automatically generated keyword ("w") and its part of speech ("POS"), number of appearances in an invalidated claim ("$n_{invalid,w}$"), number of appearances in a claim held valid ("$n_{valid,w}$"), number of times it appears at least once in a claim ("$n_{all,w}$"),[51] validity score ("Δk_w"), and frequency score ("f_w").

Keyword (w)	POS	$n_{all,w}$	$n_{invalid,w}$	$n_{valid,w}$	Validity Score (Δk_w)	Frequency Score (f_w)
output	noun, singular or mass	8	8	0	-1	0.154
part	noun, singular or mass	7	7	0	-1	0.135
price	noun, singular or mass	7	7	0	-1	0.135
processor	noun, singular or mass	7	7	0	-1	0.135
transactions	noun, plural	7	7	0	-1	0.135
input	noun, singular or mass	6	6	0	-1	0.115
group	noun, singular or mass	5	5	0	-1	0.096
property	noun, singular or mass	5	5	0	-1	0.096
describing	verb, gerund or present participle	4	4	0	-1	0.077
digital	adjective	4	4	0	-1	0.077
identifying	verb, gerund or present participle	4	4	0	-1	0.077
independent	noun, singular or mass	4	4	0	-1	0.077
less	adjective, comparative	4	4	0	-1	0.077
performed	verb, past participle	4	4	0	-1	0.077
rate	noun, singular or mass	4	4	0	-1	0.077
record	noun, singular or mass	4	4	0	-1	0.077
retrieval	noun, singular or mass	4	4	0	-1	0.077
sending	verb, gerund or present participle	4	4	0	-1	0.077
unit	noun, singular or mass	4	4	0	-1	0.077

[51] *See supra* Part III, Subpart C.

Keyword (w)	**POS**	$n_{all,w}$	$n_{vinalid,w}$	$n_{valid,w}$	**Validity Score** (Δk_w)	**Frequency Score** (f_w)
allowing	verb, gerund or present participle	3	3	0	-1	0.058
as	adverb	3	3	0	-1	0.058
asset	noun, singular or mass	3	3	0	-1	0.058
assigned	verb, past participle	3	3	0	-1	0.058
buyer	noun, singular or mass	3	3	0	-1	0.058
capture	verb, base form	3	3	0	-1	0.058
cells	noun, plural	3	3	0	-1	0.058
channel	noun, singular or mass	3	3	0	-1	0.058
color	noun, singular or mass	3	3	0	-1	0.058
comprised	verb, past tense	3	3	0	-1	0.058
comprises	noun, plural	3	3	0	-1	0.058
connected	verb, past participle	3	3	0	-1	0.058
consummating	noun, singular or mass	3	3	0	-1	0.058
content	noun, singular or mass	3	3	0	-1	0.058
corresponding	noun, singular or mass	3	3	0	-1	0.058
credit	noun, singular or mass	3	3	0	-1	0.058
debit	noun, singular or mass	3	3	0	-1	0.058
describing	noun, singular or mass	3	3	0	-1	0.058
destination	noun, singular or mass	3	3	0	-1	0.058
destinations	noun, plural	3	3	0	-1	0.058
entered	verb, past tense	3	3	0	-1	0.058
exploded	verb, past participle	3	3	0	-1	0.058
forming	verb, gerund or present participle	3	3	0	-1	0.058
function	noun, singular or mass	3	3	0	-1	0.058
goal	noun, singular or mass	3	3	0	-1	0.058
image	noun, singular or mass	3	3	0	-1	0.058
included	verb, past participle	3	3	0	-1	0.058
indicative	adjective	3	3	0	-1	0.058
input	verb, past participle	3	3	0	-1	0.058
intended	verb, past participle	3	3	0	-1	0.058
objects	noun, plural	3	3	0	-1	0.058

Keyword (w)	POS	$n_{all,w}$	$n_{invalid,w}$	$n_{valid,w}$	Validity Score (Δk_w)	Frequency Score (f_w)
pay	verb, base form	3	3	0	-1	0.058
period	noun, singular or mass	3	3	0	-1	0.058
program-controlling	adjective	3	3	0	-1	0.058
properties	noun, plural	3	3	0	-1	0.058
providing	noun, singular or mass	3	3	0	-1	0.058
purchase	noun, singular or mass	3	3	0	-1	0.058
render	noun, singular or mass	3	3	0	-1	0.058
representing	verb, gerund or present participle	3	3	0	-1	0.058
reproduction	noun, singular or mass	3	3	0	-1	0.058
rules	noun, plural	3	3	0	-1	0.058
sale	noun, singular or mass	3	3	0	-1	0.058
selected	verb, past tense	3	3	0	-1	0.058
selection	noun, singular or mass	3	3	0	-1	0.058
sent	noun, singular or mass	3	3	0	-1	0.058
separated	verb, past participle	3	3	0	-1	0.058
single	adjective	3	3	0	-1	0.058
so	adverb	3	3	0	-1	0.058
space	noun, singular or mass	3	3	0	-1	0.058
specific	adjective	3	3	0	-1	0.058
stored	verb, past tense	3	3	0	-1	0.058
telephone	noun, singular or mass	3	3	0	-1	0.058
transform	noun, singular or mass	3	3	0	-1	0.058
transformation	noun, singular or mass	3	3	0	-1	0.058
transmission	noun, singular or mass	3	3	0	-1	0.058
v	noun, singular or mass	3	3	0	-1	0.058
1	cardinal number	2	2	0	-1	0.038
according	verb, gerund or present participle	2	2	0	-1	0.038
acid	noun, singular or mass	2	2	0	-1	0.038
addition	noun, singular or mass	2	2	0	-1	0.038
adjustment	noun, singular or mass	2	2	0	-1	0.038

Keyword (w)	POS	$n_{all,w}$	$n_{invalid,w}$	$n_{valid,w}$	Validity Score (Δk_w)	Frequency Score (f_w)
and/or	noun, singular or mass	2	2	0	-1	0.038
application	noun, singular or mass	2	2	0	-1	0.038
assigning	verb, gerund or present participle	2	2	0	-1	0.038
assignment	noun, singular or mass	2	2	0	-1	0.038
Bingo	proper noun, singular	2	2	0	-1	0.038
buyers	noun, plural	2	2	0	-1	0.038
central	adjective	2	2	0	-1	0.038
column	noun, singular or mass	2	2	0	-1	0.038
columns	noun, plural	2	2	0	-1	0.038
comprises	verb, third person singular present	2	2	0	-1	0.038
computation	noun, singular or mass	2	2	0	-1	0.038
compute	verb, base form	2	2	0	-1	0.038
computerized	adjective	2	2	0	-1	0.038
computing	noun, singular or mass	2	2	0	-1	0.038
computing	verb, gerund or present participle	2	2	0	-1	0.038
condition	noun, singular or mass	2	2	0	-1	0.038
configuring	verb, gerund or present participle	2	2	0	-1	0.038
CPU	proper noun, singular	2	2	0	-1	0.038
creating	verb, gerund or present participle	2	2	0	-1	0.038
credits	noun, plural	2	2	0	-1	0.038