

Text Scaling

Slava Mikhaylov
Professor of Public Policy and Data Science

Institute for Analytics and Data Science, Department of Computer Science
Department of Government
University of Essex

Outline

Theoretical background: Treating text as data

Vector Space Model

Text Classification Models

Naive Bayes Text Classifier

Application 1

Correspondence Analysis for Text Classification

Application 2: Scaling up

United Nations General Debate

UN General Debates corpus

Corpus validation

Text as data

Text as data

- ▶ Text is ubiquitous in every aspect of our lives.
- ▶ From business, to government, to every day communication: Emails, Twitter, Facebook, news media, government reports, speeches, shareholder reports, product reviews, customer feedback, Wikileaks, Panama Papers, etc.
- ▶ Our goal is to generate insights and intelligence from these raw, unstructured (or semi-structured) data. In other words, we're in for knowledge discovery.
- ▶ In order to achieve that we need to make documents computable.

The European University Institute (EUI) in Florence, Italy, is an international postgraduate and post-doctoral teaching and research institute established by European Union member states to contribute to cultural and scientific development in the social sciences, in a European perspective.

[Contents](#) [show](#)

Departments and centres [\[edit \]](#)

Department of Economics [\[edit \]](#)

The Economics Department provides teaching and supervision to PhD students. The research activities of the Department reflect the interests of the faculty and are concentrated in micro, macro and econometrics. Weekly research seminars are given by scholars from around the world. The teaching in the doctoral programme is based on formal coursework at a level which will allow researchers to pursue academic careers in universities or to follow professional opportunities in international organisations. In their third and fourth years researchers work on their thesis projects under the guidance of their supervisor while attending research workshops and seminars.

Department of History and Civilization [\[edit \]](#)

The Department of History and Civilization (HEC) offers a programme of transnational and comparative European history. The doctoral programme studies the construction of Europe's boundaries and the diversity and complexity of experiences within them.

The department's central concerns are the interlinking of European societies since the Renaissance and the complex cultural legacies that have shaped contemporary Europe. The HEC community is also committed to exploring the place of Europe in the world through the study of empires, global processes and institutions.

The variety of research approaches and themes, as well as the broad background of its professors, enable the Department to recruit high-quality Ph.D. candidates and to host outstanding research fellows.

Department of Law [\[edit \]](#)

The Department of Law is European and international in character. It is committed to the study of law in a comparative and contextual manner, with a special focus on European and international law.

Courses and seminars are interactive, research-oriented and designed to cover the main subject areas of the Department's work. Researchers gain experience in presenting their work, and are encouraged to participate in conferences, workshops and the Department's Working Groups.

Within the department, the Academy of European Law (AEL) offers advanced-level summer courses in Human Rights Law and EU Law. It also manages research projects and runs a publications programme.

The EUI Law Department also jointly hosts, with Harvard Law School, the Summer School on Law and Logic. This summer school was launched in 2012 and is also sponsored by CIRSFID-University of Bologna (Italy), the University of Groningen (the Netherlands), the European Academy of Legal Theory, and a grant from the Erasmus Lifelong Learning Programme.

Department of Political and Social Sciences [\[edit \]](#)

The research programme of the Department of Political and Social Sciences (SPS) places emphasis on political and social change within Europe at the national, sub-national and transnational level. The research interests of the Department range across the four sub-disciplines of comparative politics, sociology, international relations, and social and political theory.

Courses in both quantitative and qualitative methods are available as options in the first and second year, while field work and data collection normally take place in the second and third year.

Robert Schuman Centre for Advanced Studies [\[edit \]](#)

The interdisciplinary centre, the Robert Schuman Centre for Advanced Studies, pursues research into the key political themes of contemporary European societies, and brings together specialists in given themes, favouring international and comparative aspects of research. The Centre has five main research projects, listed below.

European University Institute (EUI)

	
Logo of the EUI	
Type	Intergovernmental organisation
Established	1978
Endowment	€61,845,000 ^[1]
President	François Delbos
Doctoral students	Annually 140 postgraduate researchers for four years
Address	Via dei Roccettini, 9 - 50014 San Domenico di Fiesole (Italy), Florence, Italy
Campus	San Domenico di Fiesole
Website	http://www.eui.eu/

Europa 2020, la estrategia para un crecimiento inteligente, sostenible e integrador, define los objetivos que la Unión Europea debe alcanzar en el año 2020 en los ámbitos de investigación e innovación, cambio climático y energía, empleo, educación y reducción de la pobreza. También incluye siete iniciativas emblemáticas que constituyen el marco en el que la UE y las administraciones de los estados miembros pueden unir esfuerzos y reforzarse mutuamente en áreas relacionadas con las prioridades de la estrategia Europa 2020.

Para el periodo 2014-2020, la Comisión Europea ha definido un enfoque integrado de todos los fondos de la política de cohesión, mediante un marco estratégico común y con una fijación clara de prioridades y de resultados que se deben alcanzar. Asimismo, la Comisión Europea exige que la especialización inteligente sea una condición previa para las inversiones en investigación e innovación cofinanciadas con fondos europeos. Los estados miembros y las regiones deben elaborar estrategias de investigación e innovación para la especialización inteligente (*research innovation strategies for smart specialisation*, RIS3) que potencien, de acuerdo con la metodología fijada por la Comisión Europea, las especializaciones económicas y de conocimiento que se ajusten mejor a su potencial de innovación, sobre la base de los activos y las capacidades del territorio.

En este contexto, la estrategia de investigación e innovación para la especialización inteligente de Cataluña (RIS3CAT) define el marco a partir del cual el Gobierno desarrolla las actuaciones y los programas de I+D+i para el periodo 2014-2020 y da apoyo a la generación y el desarrollo de proyectos innovadores. La estrategia digital de Cataluña para el 2020 define el marco para las inversiones y las actuaciones en el ámbito de las TIC.

How to represent a document

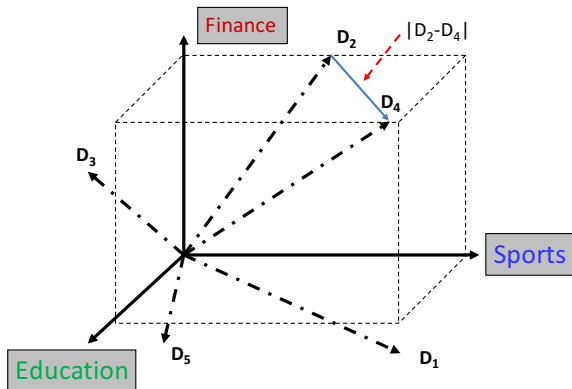
- ▶ We cannot represent by a string as it has no semantic meaning.
- ▶ We cannot represent by a list of sentences since a sentence is just like a short document (recursive definition).

Vector Space Model

- ▶ Represent documents by **concept** vectors
 - ▶ Each concept defines one dimension
 - ▶ k concepts define a high-dimensional space
 - ▶ Element of vector corresponds to concept weight: e.g.
 $d = (x_1, \dots, x_k)$, x_i is “importance” of concept i in d .
- ▶ Distance between the vectors in this concept space – relationship among documents.

An illustration of VS model

All documents are projected into this concept space



What the VS model doesn't say

- ▶ How to define/select the “basic concept”
 - ▶ Concepts are assumed to be orthogonal
- ▶ How to assign weights
 - ▶ Weights indicate how well the concept characterizes the document
- ▶ How to define the distance metric

What is a good “Basic Concept”?

- ▶ Orthogonal
 - ▶ Linearly independent basis vectors: “Non-overlapping” in meaning
 - ▶ No ambiguity
- ▶ Weights can be assigned automatically and accurately
- ▶ Existing solutions
 - ▶ Terms or N-grams, a.k.a., Bag-of-Words
 - ▶ Topics

Bag-of-Words representation

Term as the basis for vector space

- ▶ Doc1: Text mining is to identify useful information.
- ▶ Doc2: Useful information is mined from text.
- ▶ Doc3: Apple is delicious.

	text	information	identify	mining	mined	is	useful	to	from	apple	delicious
Doc1	1	1	1	1	0	1	1	1	0	0	0
Doc2	1	1	0	0	1	1	1	0	1	0	0
Doc3	0	0	0	0	0	1	0	0	0	1	1

Tokenization

Break a stream of text into meaningful units

- ▶ Tokens: words, phrases, symbols

Input: It's not straight-forward to perform so-called "tokenization."

Output(1): 'It's', 'not', 'straight-forward', 'to', 'perform', 'so-called', '"tokenization".'

Output(2): 'It', "'", 's', 'not', 'straight', '-', 'forward', 'to', 'perform', 'so', '-', 'called', '"', 'tokenization', ',', "'", ''

- ▶ Definition depends on language, corpus, or even context.
- ▶ We usually approach tokenization through regular expressions or use of statistical methods (Apache OpenNLP, Stanford NLP Parser).

Bag-of-Words representation

	text	information	identify	mining	mined	is	useful	to	from	apple	delicious
Doc1	1	1	1	1	0	1	1	1	0	0	0
Doc2	1	1	0	0	1	1	1	0	1	0	0
Doc3	0	0	0	0	0	1	0	0	0	1	1

- ▶ Assumption: Words are independent from each other
- ▶ Pros: Simple
- ▶ Cons: Basis vectors are clearly not linearly independent!
Grammar and order are missing.
- ▶ The most frequently used document representation (also used in image, speech, gene sequence).

Bag-of-Words with N-grams

- ▶ N-grams: a contiguous sequence of n tokens from a given piece of text
 - ▶ E.g., 'Text mining is to identify useful information.'
 - ▶ Bigrams: 'text_mining', 'mining_is', 'is_to', 'to_identify', 'identify_useful', 'useful_information', 'information_.'
- ▶ Pros: capture local dependency and order
- ▶ Cons: makes computation more complicated due to term sparseness.

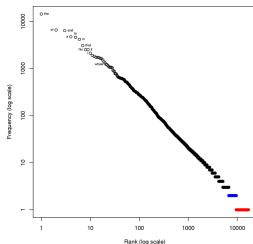
Automatic document representation

Represent a document with all the occurring words

- ▶ Pros: Preserve all information in the text (hopefully). Also, it's fully automatic.
- ▶ Cons: vocabulary gap – cars vs car, talk vs talking. Leads to large storage demands on computing.
- ▶ Solution: Construct controlled vocabulary.

Statistical properties of language

- ▶ Rank-frequency plot for words in the novel Moby-Dick.
- ▶ About 44% of the distinct set of words in this novel, such as “matrimonial”, occur only once, so called hapax legomena (red).
- ▶ About 17%, such as “dexterity”, appear twice (so-called dis legomena, in blue).
- ▶ Zipf’s law predicts that the words in this plot should approximately fit a straight line – frequency of any word is inversely proportional to its rank in the frequency table.



Controlled vocabulary

- ▶ Head words take large portion of occurrences, but they are semantically meaningless. E.g., the, a, an, we, do, to.
- ▶ Tail words take major portion of vocabulary, but they rarely occur in documents. E.g., matrimonial in Moby-Dick.
- ▶ The rest is most representative, and should be included in the controlled vocabulary.

Normalization

- ▶ Convert different forms of a word to a normalized form in the vocabulary. For example, U.S.A. to USA, St. Louis to Saint Louis.
- ▶ Solution
 - ▶ Rule-based
 - ▶ Delete periods and hyphens,
 - ▶ All in lower cases.
 - ▶ Dictionary-based
 - ▶ Construct equivalent class: Car to “automobile, vehicle”; Mobile phone to “cellphone”.

Stemming

Lemmatization refers to the algorithmic process of converting words to their lemma forms.

stemming the process for reducing inflected (or sometimes derived) words to their stem, base or root form. Different from *lemmatization* in that stemmers operate on single words without knowledge of the context.

both convert the morphological variants into stem or root terms

example: **produc** from
production, producer, produce, produces,
produced

Stopwords

Useless words for document analysis

- ▶ Not all words are informative.
- ▶ Remove such words to reduce vocabulary size.
- ▶ No universal definition.
- ▶ We risk breaking the original meaning and structure of text.

Common English stop words

a, able, about, across, after, all, almost, also, am, among, an, and, any, are, as, at, be, because, been, but, by, can, cannot, could, dear, did, do, does, either, else, ever, every, for, from, get, got, had, has, have, he, her, hers, him, his, how, however, I, if, in, into, is, it, its, just, least, let, like, likely, may, me, might, most, must, my, neither, no, nor, not, of, off, often, on, only, or, other, our, own, rather, said, say, says, she, should, since, so, some, than, that, the, their, them, then, there, these, they, this, tis, to, too, twas, us, wants, was, we, were, what, when, where, which, while, who, whom, why, will, with, would, yet, you, your

- But no list should be considered universal

A more comprehensive list of stop words

a's, able, about, above, according, accordingly, across, actually, after, afterwards, again, against, ain't, all, allow, allows, almost, alone, along, already, also, although, always, am, among, amongst, an, and, another, any, anybody, anyhow, anyone, anything, anyway, anyways, anywhere, apart, appear, appreciate, appropriate, are, aren't, around, as, aside, ask, asking, associated, at, available, away, awfully, be, became, because, become, becomes, becoming, been, before, beforehand, behind, being, believe, below, beside, besides, best, better, between, beyond, both, brief, but, by, c'mon, c's, came, can, can't, cannot, cant, cause, causes, certain, certainly, changes, clearly, co, com, come, comes, concerning, consequently, consider, considering, contain, containing, contains, corresponding, could, couldn't, course, currently, definitely, described, despite, did, didn't, different, do, does, doesn't, doing, don't, done, down, downwards, during, each, edu, eg, eight, either, else, elsewhere, enough, entirely, especially, et, etc, even, ever, every, everybody, everyone, everything, everywhere, ex, exactly, example, except, far, few, fifth, first, five, followed, following, follows, for, former, formerly, forth, four, from, further, furthermore, get, gets, getting, given, gives, go, goes, going, gone, got, gotten, greetings, had, hadn't, happens, hardly, has, hasn't, have, haven't, having, he, he's, hello, help, hence, her, here, here's, hereafter, hereby, herein, hereupon, hers, herself, hi, him, himself, his, hither, hopefully, how, howbeit, however, i'd, i'll, i'm, i've, ie, if, ignored, immediate, in, inasmuch, inc, indeed, indicate, indicated, indicates, inner, insofar, instead, into, inward, is, isn't, it, it'd, it'll, it's, its, itself, just, keep, keeps, kept, know, knows, known, last, lately, later, latter, latterly, least, less, lest, let, let's, like, liked, likely, little, look, looking, looks, ltd, mainly, many, may, maybe, me, mean, meanwhile, merely, might, more, moreover, most, mostly, much, must, my, myself, name, namely, nd, near, nearly, necessary, need, needs, neither, never, nevertheless, new, next, nine, no, nobody, non, none, noone, nor, normally, not, nothing, novel, now, nowhere, obviously, of, off, often, oh, ok, okay, old, on, once, one, ones, only, onto, or, other, others, otherwise, ought, our, ours, ourselves, out, outside, over, overall, own, particular, particularly, per, perhaps, placed, please, plus, possible, presumably, probably, provides, que, quite, qv, rather, rd, re, really, reasonably, regarding, regardless, regards, relatively, respectively, right, said, same, saw, say, saying, says, second, secondly, see, seeing, seem, seemed, seeming, seems, seen, self, selves, sensible, sent, serious, seriously, seven, several, shall, she, should, shouldn't, since, six, so, some, somebody, somehow, someone, something, sometime, sometimes, somewhat, somewhere, soon, sorry, specified, specify, specifying, still, sub, such, sup, sure, t's, take, taken, tell, tends, th, than, thank, thanks, thanx, that, that's, thats, the, their, theirs, them, themselves, then, thence, there, there's, thereafter, thereby, therefore, therein, theres, thereupon, these, they, they'd, they'll, they're, they've, think, third, this, thorough, thoroughly, those, though, three, through, throughout, thru, thus, to, together, too, took, toward, towards, tried, tries, truly, try, trying, twice, two, un, under, unfortunately, unless, unlikely, until, unto, up, upon, us, use, used, useful, uses, using, usually, value, various, very, via, viz, vs, want, wants, was, wasn't, way, we, we'd, we'll, we're, we've, welcome, well, went, were, weren't, what, what's, whatever, when, whence, whenever, where, where's, whereafter, whereas, whereby, wherein, whereupon, wherever, whether, which, while, whither, who, who's, whoever, whole, whom, whose, why, will, willing, wish, with, within, without, won't, wonder, would, would, wouldn't, yes, yet, you, you'd, you'll, you're, you've, your, yours, yourself, yourselves, zero

Constructing a VSM representation

1. Tokenization;
2. Stemming/normalization;
3. N-gram construction;
4. Stopword/controlled vocabulary filtering.

Assigning weights

- ▶ Corpus-wise: some terms carry more information about the document content.
- ▶ Document-wise: not all terms are equally important.
- ▶ Two basic heuristics: **TF** (Term Frequency) is the Within-doc-frequency; **IDF** (Inverse Document Frequency).

Term Frequency normalization

- ▶ Idea: a term is more important if it occurs more frequently in a document.
- ▶ Two views of document length:
 - ▶ A doc is long because it is verbose.
 - ▶ A doc is long because it has more content.
- ▶ Raw TF is inaccurate
 - ▶ Document length variation.
 - ▶ “Repeated occurrences” are less informative than the “first occurrence.”
 - ▶ Semantic information does not increase proportionally with number of term occurrences.
- ▶ Generally penalize long document, but avoid over-penalizing – pivoted length normalization

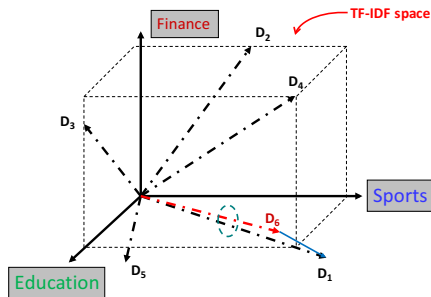
Document Frequency normalization

- ▶ Idea: a term is more discriminative if it occurs only in fewer documents.
- ▶ Inverse document frequency: Assign higher weights to the rare terms.
- ▶ This is a corpus-specific property. Independent of a single document.

TF-IDF weighting

- ▶ Combining TF and IDF
 - ▶ Common in document \rightarrow high tf \rightarrow high weight;
 - ▶ Rare in collection \rightarrow high idf \rightarrow high weight;
 - ▶ $w(t, d) = TF(t, d) \times IDF(t)$
- ▶ Most well-known document representation schema in information retrieval.

Defining a similarity metric



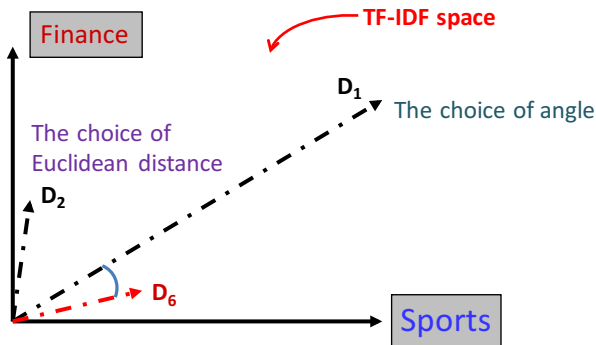
We can use Euclidean distance as a similarity metric:

$$\text{dist}(d_i, d_j) = \sqrt{\sum_{t \in V} [tf(t, d_i)idf(t) - tf(t, d_j)idf(t)]^2}$$

- ▶ Longer documents will be penalized by the extra words;
- ▶ We care more about how these two vectors overlap.

From distance to angle

- ▶ Focusing on the angle tells us how vectors overlap.
- ▶ Cosine similarity – projection of one vector onto another.



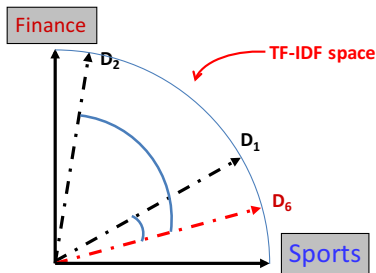
Cosine similarity

- ▶ Angle between two vectors

$$\text{cosine}(d_i, d_j) = \frac{v_{d_i} \times v_{d_j}}{|v_{d_i}|_2 \times |v_{d_j}|_2}$$

where v_{d_i} is the TF-IDF vector and $\frac{v_{d_j}}{|v_{d_j}|_2}$ is unit vector.

- ▶ Documents are normalized by length.



Advantages of VS model

- ▶ Empirically effective.
- ▶ Intuitive.
- ▶ Easy to implement.
- ▶ Well-studied/mostly evaluated.
- ▶ The Smart system developed at Cornell (1960-1999) but still widely used today.

Disadvantages of VS model

- ▶ We assume term independence.
- ▶ Lack of “predictive adequacy” due to arbitrary term weighting and arbitrary similarity measure.
- ▶ Lots of parameter tuning!

What you should take out

- ▶ Basic ideas of vector space model.
- ▶ Procedures of constructing VS representation for a document.
- ▶ Two important heuristics in bag-of-words representation: TF and IDF.
- ▶ Similarity metric for VS model.

Text Classification Models

Text Classification

Text classification (TC) is applicable to a wide range of problems:

- ▶ Detecting document encoding;
- ▶ Spam filtering;
- ▶ Sentiment detection;
- ▶ Document sorting in smart email folders.

Text Classification

- ▶ Text has traditionally been classified manually (e.g. a library catalogue).
- ▶ However, this is not scalable.
- ▶ In machine learning, the decision criterion of the text classifier is learned automatically from training data.
- ▶ In such statistical text classification we require a number of good example documents (training documents) for each class.
- ▶ Training documents are manually labeled (annotated with class label).

The text classification problem

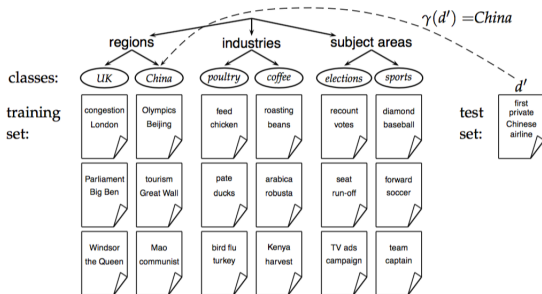
- ▶ In TC we are given a description $d \in X$ of a document, where X is the **document space**; and a fixed set of **classes** $C = \{c_1, c_2, \dots, c_J\}$.
- ▶ Classes are also called **categories** or **labels**.
- ▶ The document space X is a high-dimensional space, and the classes are human defined for the need of an application.
- ▶ We are also given a **training set** D of labeled documents $\langle d, c \rangle \in X \times C$.
- ▶ For example:

$\langle d, c \rangle = \langle \textit{Beijing joins the World Trade Organization}, \textit{China} \rangle$

for the one-sentence document *Beijing joins the World Trade Organization* and the class (or label) *China*.

The text classification problem

- ▶ Function that maps documents to classes is a classifier
 $\gamma : X \rightarrow C$.
- ▶ We are using supervised learning here since a human needs to define the classes and label training documents.
- ▶ We apply the classifier to the test set, whose class is unknown. Our goal is to achieve high accuracy on the test data.
- ▶ In order to achieve that we have to make the assumption that training data and test data are similar or from **the same distribution**.



► **Figure 13.1** Classes, training set, and test set in text classification .

Naive Bayes text classifier

- ▶ Multinomial Naive Bayes (multinomial NB) is a probabilistic learning method, an extension of what we covered when we discussed classification models.
- ▶ The probability of a document d being in class c is computed as

$$P(c|d) \propto P(c) \prod_{1 \leq k \leq n_d} P(t_k|c)$$

where $P(t_k|c)$ is the conditional probability of term t_k occurring in a document of class c .

- ▶ $P(t_k|c)$ can be interpreted as a measure of how much evidence t_k contributes that c is the correct class.
- ▶ $P(c)$ is the prior probability of a document occurring in class c .
- ▶ $\langle t_1, t_2, \dots, t_{n_d} \rangle$ are the tokens in d that are part of the vocabulary we use for classification and n_d is the number of such tokens in d .

Maximum a posteriori class

- ▶ Our goal is to find the **best** class for the document.
- ▶ The best class in NB classification is the most likely or **maximum a posteriori** (MAP) class c_{map} :

$$c_{map} = \underset{c \in C}{\operatorname{argmax}} \hat{P}(c|d) = \underset{c \in C}{\operatorname{argmax}} \hat{P}(c) \prod_{1 \leq k \leq n_d} \hat{P}(t_k|c).$$

Underflow

- ▶ In the function above, many conditional probabilities are multiplied, one for each position $1 \leq k \leq n_d$.
- ▶ This can result in a floating point underflow (the result of a calculation is a smaller number than the computer can actually store in memory).
- ▶ Thus it's better to add logarithms of probabilities instead of multiplying probabilities.
- ▶ The class with the highest log probability score is still the most probable

$$\log(xy) = \log(x) + \log(y)$$

and the logarithm function is monotonic.

- ▶ NB maximization is then

$$c_{map} = \underset{c \in C}{argmax} [\log \hat{P}(c) + \sum_{1 \leq k \leq n_d} \log \hat{P}(t_k | c)].$$

Interpreting NB

- ▶ Each conditional parameter $\log \hat{P}(t_k|c)$ is a weight that indicates how good an indicator t_k is for c .
- ▶ The prior $\log \hat{P}(c)$ is a weight that indicates the relative frequency of c .
- ▶ More frequent classes are more likely to be the correct class than infrequent classes.
- ▶ The sum of log prior and term weights is then a measure of how much evidence there is for the document being in the class.
- ▶ c_{map} equation selects the class for which we have the most evidence.

Estimating NB model

- ▶ We can use MLE to estimate the parameters in the model. Here it's the relative frequency and corresponds to the most likely value of each parameter given the training data.
- ▶ For the priors this is:

$$\hat{P}(c) = \frac{N_c}{N},$$

where N_c is the number of documents in class c and N is the total number of documents.

- ▶ Conditional probability

$$\hat{P}(t|c) = \frac{T_{ct}}{\sum_{t' \in V} T_{ct'}},$$

where T_{ct} is the number of occurrences of t in training documents from class c , including multiple occurrences of a term in a document.

Positional independence assumption

- ▶ Here we made the **positional independence assumption**: T_{ct} is a count of occurrences in all positions k in the documents in the training set.
- ▶ We do not compute different estimates for different positions and, for example, if a word occurs twice in a document, in positions k_1 and k_2 , then $\hat{P}(t_{k_1}|c) = \hat{P}(t_{k_2}|c)$.

Zero probabilities in MLE

- ▶ The problem with the MLE estimate is that it is zero for a termclass combination that did not occur in the training data.
- ▶ If the term WTO in the training data only occurred in China documents, then the MLE estimates for the other classes, for example UK, will be zero:

$$\hat{P}(WTO|UK) = 0.$$

- ▶ Now, the one-sentence document *Britain is a member of the WTO* will get a conditional probability of zero for UK because we are multiplying the conditional probabilities for all terms in the NB equation.
- ▶ We should have a high probability for the UK class because the term Britain occurs.
- ▶ Zero probability for WTO remains regardless of the weight of evidence for the UK from other features.

Sparseness

- ▶ The estimate is 0 because of **sparseness**.
- ▶ The training data is not large enough to capture the frequency of rare events adequately.
- ▶ To eliminate zeros we use **add-one** or **Laplace smoothing**:

$$\hat{P}(t|c) = \frac{T_{ct} + 1}{\sum_{t' \in V} (T_{ct' + 1})} = \frac{T_{ct} + 1}{(\sum_{t' \in V} T_{ct'1}) + B'},$$

where $B = |V|$ is the number of terms in the vocabulary.

Properties of NB

- ▶ Class membership of a document is decided based on the class with the maximum a posteriori probability computed as:

$$c_{map} = \underset{c \in \mathcal{C}}{\operatorname{argmax}} P(c|d) = \underset{c \in \mathcal{C}}{\operatorname{argmax}} \frac{P(d|c)P(c)}{P(d)} = \underset{c \in \mathcal{C}}{\operatorname{argmax}} P(d|c)P(c),$$

where to get the last equality we apply Bayes' rule and drop the denominator because $P(d)$ is the same for all classes and does not affect the argmax.

- ▶ The last equality can be interpreted as a description of the generative process we assume in Bayesian text classification.

Text generative process

1. Choose class c with probability $P(c)$, where C is a random variable.
2. Generate the document given the class, corresponding to the conditional distribution $P(d|c)$:

$$P(d|c) = P(\langle t_1, \dots, t_k, \dots, t_{n_d} \rangle | c)$$

where $\langle t_1, \dots, t_{n_d} \rangle$ is the sequence of terms as it occurs in d (minus terms that were excluded from the vocabulary).

NB CIA

- ▶ The number of parameters in such a model would be huge and we won't be able to perform text classification directly.
- ▶ Therefore, we introduce the Naive Bayes **conditional independence assumption** – attribute values are independent of each other given the class:

$$P(d|c) = P(\langle t_1, \dots, t_k, \dots, t_{n_d} \rangle | c) = \prod_{1 \leq k \leq n_d} P(X_k = t_k | c)$$

- ▶ X_k is the random variable for position k in the document.
- ▶ X_k takes as values terms from the vocabulary.
- ▶ $P(X_k = t | c)$ is the probability that in a document of class c the term t will occur in position k .

Positional independence

- ▶ Even with the NB CIA assumption there are still too many parameters to estimate.
- ▶ The position of a term in a document by itself does not carry information about the class.
- ▶ There is a difference between **China sues France** and **France sues China**. But the occurrence of China in position 1 versus position 3 of the document is not useful in NB classification because we look at each term separately.
- ▶ Also, estimating different set of parameters for each k could lead to problems due to data sparseness.

Positional independence assumption

- ▶ Hence we make the **Positional Independence Assumption** (PIA) – conditional probabilities for a term are the same independent of position in the document:

$$P(X_{k_1} = t|c) = P(X_{k_2} = t|c)$$

for all positions k_1, k_2 , terms t and classes c .

- ▶ This leads to a single distribution of terms valid for all positions k_i .
- ▶ PIA is equivalent to adopting the bag of words model.

NB as a good classifier

- ▶ Naive Bayes gets its name from CIA and PIA being very naive models of natural language.
- ▶ Due to such poor models of natural language NB probability estimates are of low quality.
- ▶ However, NB classification decisions are surprisingly good.
- ▶ NB classification decision is based on which class gets the highest score. It does not matter how accurate the estimates are.
- ▶ Correct estimation implies accurate prediction, but accurate prediction does not imply correct estimation.
- ▶ NB classifiers estimate badly, but often classify well.

Benefits of NB

- ▶ NB is robust to noise features and **concept drift**.
- ▶ Other classifiers (e.g. kNN) may be more accurate but also more tuned to idiosyncratic properties of a particular time period.
- ▶ They will suffer if documents in the following time period have slightly different properties.

Benefits of NB

- ▶ NB is extremely efficient with training and classification done in one pass over the data.
- ▶ NB through its combination of efficiency and good accuracy is often used as a baseline in text classification.
- ▶ Where CIA and PIA hold it can be shown that NB is an **optimal classifier** (minimal test error rate) for text.

Using NB

We often prefer NB when:

- ▶ marginal increase in accuracy is not that important for the text classification application at hand,
- ▶ a very large amount of training data is available and there is more to be gained from training on a lot of data than using a better classifier on a smaller training set, or
- ▶ its robustness to concept drift can be exploited.

$tf_{t_i,d}$ formulation of NB

- ▶ A useful alternative formalization of NB treats each document d as an M -dimensional vector of counts $\langle tf_{t_1,d}, \dots, tf_{t_M,d} \rangle$ where $tf_{t_i,d}$ is the term frequency of t_i in d .
- ▶ Then $P(d|c)$ is computed as follows:

$$P(d|c) = P(\langle tf_{t_1,d}, \dots, tf_{t_M,d} \rangle | c) \propto \prod_{1 \leq i \leq M} P(X = t_i | c)^{tf_{t_i,d}}$$

- ▶ This is equivalent to the sequence model in our first formulation of multinomial Naive Bayes earlier as $P(X = t_i | c)^{tf_{t_i,d}} = 1$ for terms that do not occur in d ($tf_{t_i,d} = 0$) and a term that occurs $tf_{t_i,d} \geq 1$ times will contribute $tf_{t_i,d}$ factors in both formulations of multinomial NB.

Application 1: Analyzing intra-cabinet politics with NB

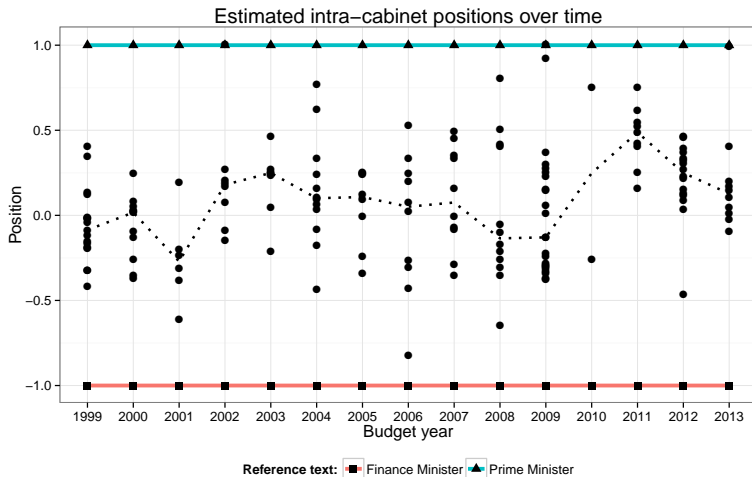
Analyzing intra-cabinet politics with NB

- ▶ It is difficult to assess intra-cabinet politics. But it's important for many political science models to have estimates of ministers' positions on various policies.
- ▶ In Ireland most ministers speak during the budget debates. We can use these speeches to categorize ministers as belonging to one of the two classes (e.g. PM vs FM).
- ▶ We can use probabilities of belonging to each class (scores) as measures of positions on a policy dimension defined by the training documents (PM speeches and FM speeches).
- ▶ In political science jargon this is often called **ideological scaling** or **position scaling**.

NB with political texts

- ▶ One “native” implementation of the Naive Bayes classifier is called **Wordscore** presented in Laver, Michael, Kenneth Benoit and John Garry. 2003. “Extracting Policy Positions from Political Texts Using Words as Data.” *American Political Science Review* 97: 311-331.
- ▶ Computation implemented in R `quanteda` package:
<https://github.com/kbenoit/quanteda>.

Applying NB to budget debates in Ireland



Alexander Herzog and Slava Mikhaylov "Intra-Cabinet Politics and Fiscal Governance in the Times of Austerity" (working paper).

Correspondence Analysis for Text Classification

Correspondence Analysis for text classification

- ▶ Will Lowe (2008) showed that main text scaling models popular in political science (including Wordscore) can be formalized as special cases of correspondence analysis.
 - ▶ Lowe, William (2008). "Understanding Wordscores." *Political Analysis* 16(4): 356-371.
- ▶ The assumptions of correspondence analysis fit nicely with natural language properties and model outputs interesting for political science.

Correspondence Analysis

- ▶ CA is an SVD technique, just like PCA.
- ▶ It replaces the Euclidean metric with Chi-square metric, and weights rows (“sites”) and columns (“species”) by their totals.
- ▶ The differences may sound small and technical, but in practice, the difference is huge.

Correspondence Analysis

Invented independently numerous times:

1. **Correspondence Analysis:** Weighted Principal Components with Chi-squared metric.
2. **Optimal Scaling:** Find site and species scores so that
 - ▶ all species occurring in one site are as similar as possible, but
 - ▶ species of different sites are as different as possible, and
 - ▶ sites are dispersed as widely as possible relative to species scores.
3. **Reciprocal Averaging:** Species scores are weighted averages of site scores, and simultaneously, site scores are weighted averages of species scores.

Necessary vocabulary from ecological analysis

- ▶ **Species response curve** – a graphical portrayal of the abundance of a species as a function of an environmental gradient.
- ▶ **Abundance** – a measure of the amount of a species in a sample (e.g. density, frequency, territorial area).
- ▶ **Environmental gradient** – a spatially varying aspect of the environment which is expected to be related to species composition.

Chi-squared metric

- ▶ Metric is a 'yardstick' to measure dissimilarities among points.
- ▶ PCA uses constant ('Euclidean') metric, but CA uses expected abundances as a metric.
- ▶ Expected abundances from marginal totals: Exactly like in χ^2 analysis of contingency tables.
- ▶ *Species profile* is the average proportion of columns in the data, and *site profile* is the average proportion of rows.
- ▶ All sites should have all species in the same proportions.
- ▶ Chi-squared distance is the difference between expected profile and real abundance distributions – both for the species and the sites.

CA vs PCA

- ▶ Practical application of CA is often driven by an identified problem of PCA – “**horseshoe effect**” that makes it unsuitable for use in many fields (e.g. ecology, text mining).
- ▶ The problem is caused by the fact that species often have **unimodal** species response curves along environmental gradients.
- ▶ In the case of species response curves, a unimodal distribution means the species has one optimal environmental condition.
- ▶ If any aspect of the environment is greater or lesser than this optimum, the species will perform more poorly (i.e. it will have a lesser abundance).
- ▶ Some techniques (such as CA) perform best when species have **unimodal distributions**, others (such as PCA) perform better when species have **monotonic distributions** along gradients (i.e. the species either increase or decrease, but not both, as a function of environmental factors).

CA vs PCA

- ▶ PCA assumes that species are linearly (or at least monotonically) related to each other and to gradients.
- ▶ PCA fails because it represents site occurrences in species space (i.e. emphasis is on uncovering trends in species abundance).
- ▶ CA represents species **and** sites as occurring in a postulated environment space (often called ordination space).
- ▶ CA assumes that species have unimodal species response curves. A species is located in that location of space where it is most abundant.

Application 2: Scaling up

Motivation

- ▶ Understanding government preferences is central to the study of international relations and comparative politics.
- ▶ Standard operationalization and measurement:
 1. Alliance commitments (Signorino and Ritter 1999).
 - ▶ Reflect security environment and geographic proximity.
 - ▶ Informal alliances.
 2. Voting patterns in the UN General Assembly (Russett 1966; Voeten 2000, 2004, 2013; Bailey et al. 2013).
 - ▶ Dynamic assessment of preferences: changes in preferences vs changes in UN agenda.
 - ▶ 75% of resolutions adopted without a vote.
 - ▶ Important votes.
 - ▶ Strategic voting (and non-voting).

Contribution

- ▶ An alternative approach based on the UN General Debate speeches.
- ▶ “State-of-the-World” addresses.
- ▶ Country perspectives on key international and domestic issues.
- ▶ Relative emphasis and prioritization of these issues.
- ▶ Government preferences on meaningful dimensions.

The General Debate of the United Nations



“Speeches at the General Debate are interesting because they flesh out national policies – what states think. The General Debate is the one place where states can speak their mind; it reflects the issues that states consider important.”

Deputy Head of Finnish Mission to the UN.

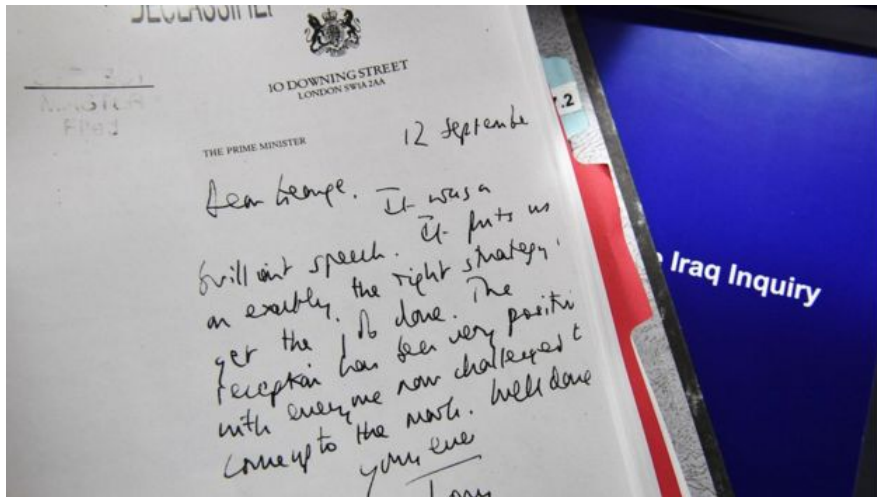
“These speeches are the most sovereign thing that a country does as a member of the UN.”

Spokesperson for the German Mission to the UN.

UNGD 2002: Strategy and politics



The Case for War in Iraq



UN General Debates corpus

- ▶ All speeches from 1970 to 2014.
- ▶ United Nations Dag Hammarskjold Library: English version.
- ▶ 7,310 statements.
- ▶ UN membership: 70 (1970) and 193 (2014).
- ▶ On average 123 sentences, 945 *unique forms*, and 3,248 *individual* words each.

The meeting was suspended at 11.25 a.m. and resumed at 11.50 a.m.

ADDRESS BY MR. GEORGE BUSH, PRESIDENT OF THE UNITED STATES OF AMERICA

The PRESIDENT: The Assembly will now hear an address by the President of the United States of America.





























Mr. George Bush, President of the United States of America, was escorted into the General Assembly Hall.

The PRESIDENT: On behalf of the General Assembly I have the honour to welcome to the United Nations the President of the United States of America, His Excellency Mr. George Bush, and to invite him to address the General Assembly.

President BUSH: I am honoured to address the General Assembly today at the beginning of its forty-fourth session. I should like to congratulate Joseph Garba of Nigeria, a distinguished diplomat, on his election as President of this session of the General Assembly, and I wish him success in his presidency.

I feel a great personal pleasure on this occasion, for this is a homecoming

Corpus overview

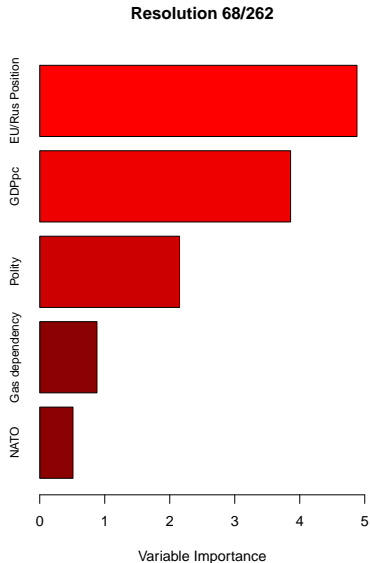
topic	1970—2014	top words	proportion of corpus	
Topic 5		international nations united economic development efforts countries peace world security cooperation general states community assembly		12.1%
Topic 8		international organization charter states peoples general time countries community situation principles fact assembly make respect		7.4%
Topic 6		nations united world time years war today year people make future great work past ago		5.7%
Topic 2		united nations rights human security council states international nuclear weapons member organization treaty important work		5.5%
Topic 17		issues international nations assembly peace community global organization security process continue major remain remains states		5.5%
Topic 10		south africa international nations united people countries economic namibia independence world peace session apartheid situation		5.2%
Topic 20		nations united countries world session government assembly conference co-operation economic general delegation international africa development		4.9%
Topic 26		co-operation countries efforts international community problems political hope negotiations peace conference progress government parties development		4.9%
Topic 12		world political international economic system order interests relations national common nations power social development problems		4.4%
Topic 11		countries economic developing development world trade developed international resources cent growth economy financial debt food		4.0%
Topic 13		arab peace international security israel palestinian people region states resolutions israeli east middle iraq council		3.9%
Topic 4		human people world rights peace freedom life country today justice nations peoples democracy children dignity		3.5%
Topic 15		development poverty world human countries united global cooperation international general commitment council terrorism assembly security		3.5%
Topic 14		development global climate change goals security agenda support crisis energy efforts poverty commitment food world		3.4%

More than talk?



Ukraine Territorial Integrity

- ▶ Vote on Resolution 68/262
- ▶ Random forest estimation
- ▶ Out-of-bag test error rate 7.78%
- ▶ Variable importance – aggregate improvement in the optimization objective for each predictor.



Other methods

Document-Term Matrix

Within the VSM approach computation in R (and most other software packages) requires creation of document-term matrix (DTM) also known as document-feature matrix:

dtm.2002	far	stood	two	tower	symbol	freedom	prosper	progress	halfway	around	globe	magnific	buddha	repres	cultur	toler	nation	rich	histori
AFG_57_2002	1	2	3	1	2	2	5	1	1	1	1	1	1	5	3	4	17	2	3
AGO_57_2002	0	0	1	0	0	0	0	1	0	0	0	0	0	0	0	0	10	0	0
ALB_57_2002	0	0	1	0	0	1	0	0	0	0	0	0	0	1	2	1	19	0	0
AND_57_2002	0	0	2	6	1	0	0	0	0	0	0	0	0	0	2	6	9	0	5
ARE_57_2002	0	0	3	0	0	0	0	0	0	0	0	0	0	1	2	0	16	0	0
ARG_57_2002	0	0	0	0	0	0	1	1	0	0	0	0	0	2	0	0	13	0	0
ARM_57_2002	0	0	3	0	0	0	1	3	0	1	0	0	0	1	0	0	8	0	2
ATG_57_2002	0	0	0	0	0	0	0	2	0	0	2	0	0	2	0	0	11	0	0
AUS_57_2002	0	0	1	0	0	0	1	0	0	0	0	0	0	1	0	0	29	0	0
AUT_57_2002	0	0	3	0	0	2	0	1	0	0	0	0	0	0	3	1	15	0	0
AZE_57_2002	3	0	0	0	0	0	0	0	0	1	0	0	0	3	2	1	13	0	1
BDI_57_2002	1	0	1	0	0	0	0	0	0	0	0	0	0	1	0	0	7	1	0
BEL_57_2002	0	1	1	0	0	2	2	3	0	0	0	0	0	0	0	2	7	1	1
BEN_57_2002	0	0	0	0	0	0	0	2	0	0	0	0	0	1	0	0	19	0	0
BFA_57_2002	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	13	0	0
BGD_57_2002	0	0	3	0	0	1	0	1	0	0	0	0	0	1	2	1	5	1	1
BGR_57_2002	1	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	9	0	0
BHR_57_2002	0	1	0	0	0	0	3	1	0	0	0	0	0	0	4	1	20	0	0
BHS_57_2002	0	0	0	0	0	1	1	2	0	1	0	0	0	2	1	0	16	0	0
BIH_57_2002	1	0	1	0	0	0	1	3	0	0	0	0	0	4	1	3	7	0	0
BLR_57_2002	0	0	6	0	0	0	0	3	0	0	0	0	0	0	0	0	18	0	0
BLZ_57_2002	0	0	4	0	0	0	0	0	0	0	0	0	0	1	0	0	19	0	1
BOL_57_2002	0	0	1	0	0	0	1	1	0	0	0	0	0	1	0	1	7	2	2

Showing 1 to 23 of 188 entries

Other text classification models

Once you have a DTM you can apply any standard supervised and unsupervised learning model. For example,

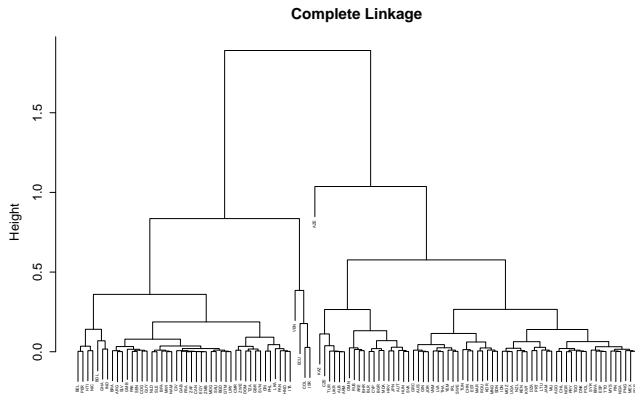
- ▶ kNN;
- ▶ Logistic regression;
- ▶ Support Vector Machines (SVMs);
- ▶ Random Forest;
- ▶ PCA/CA;
- ▶ Clustering.

More in *An Introduction to Information Retrieval*, Chapters 14-15.

Text clustering

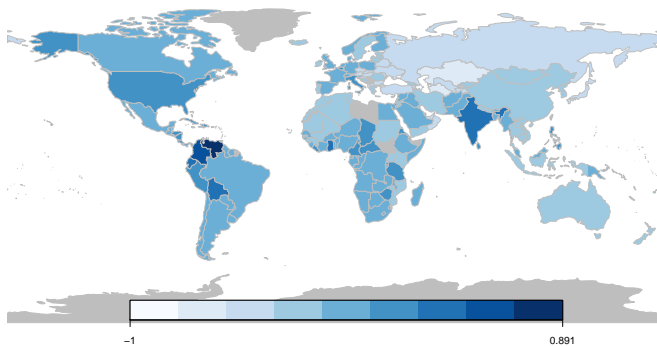
- ▶ In text clustering we are interested in identifying the clustering structure of a corpus of text documents and assigning documents to the identified cluster(s).
- ▶ We can apply to DTM standard clustering algorithms like k-means clustering (example of centroid-based clustering) and hierarchical clustering (also known as connectivity-based clustering).
- ▶ You will need to transform DTM to “distances” between documents.

Hierarchical clustering



- ▶ Complete linkage hierarchical clustering.
- ▶ Scores on the second dimension of correspondence analysis of UN corpus for 2002.

Correspondence analysis Dimension 2: 2002



What you need to take out

- ▶ Vector space model allows us to treat text as data.
- ▶ With VSM we can apply any supervised and unsupervised learning tool we've learnt so far to discover knowledge and generate insights from text.