Building a topical representation of words

Done by Andrei Constantinecu At the University of Pennsylvania – August to December 2017

Under the supervision of Professor Ani Nenkova

1. Topical representation of words

We seek to represent each word as a vector whose dimensions represent the strength of association of a word to a specific topics. Let us say we have 12 topics : Arts - Sports – US – Style – Health- Travel - Business – World - Movies – Technology – Education – Science. The first dimension in this representation of the word will reflect how associated the word is to the topic Arts, the second to the topic sport, and so on.

From this idea, one might expect to obtain the following kind of representations:
Painter = [10 0 0 0 0 0 0 0 0 0 0 0 0]
Baseball = [0 10 0 0 0 0 0 0 0 0 0 0]
Where the word ‘painter’ would be havily associated to the topic ‘Arts’, and the word ‘Baseball’ to the topic ‘Sports’.

1.1 The Data and method

To compute topical scores for each word and for each topic, we will use a New-York Times annotated corpus, and we will contrast distributions of words between a topical corpus and a reference corpus. This corpus comprises all New-York times articles published between 1987 and 2007. The articles are annotated with several pieces of metadata such as the section in which they appeared, or descriptors specifying in more detail what an article is about. For each topic, we therefore have a topical sample of articles, i.e. articles being published in the same section of the newspaper, or articles being described with the same descriptor. As a reference corpus, we ran through the entire New-York-Times corpus and computed frequencies for each word.

1.2 Five techniques to compare differences in distributions

1- Log odds: We compute a score for each word as follows:
Score_log_odds(word) = log [(frequency topic) / (frequency reference)]
This technique is equivalent to computing the Point-wise mutual information between a word and a topic.

2-Signature :This technique is based on a statistical test.
If we set H1 as the hypothesis that the distribution of the word in the topic corpus is not different from that of the reference corpus, and H2 as the hypothesis that the two are different, Lin et al. showed that:

$$-2log\frac{L(H_1)}{L(H_2)}$$

follows a chi square distribution. The higher the chi-square statistic the more likely the two distributions are different, the more likely the word is a topical word.

3- Regression delta: We compute a score by devising a regression task. If we consider a regression task where the aim is to predict the frequency of a word in the whole corpus given its frequency in a biased topical corpus, one might imagine that the prediction of topical words will overshoot.Thus we can compute a score for each word as follows:

Score_regression(word) = Prediction(word) – Reference(word)

4 - Bag of words: The ideas is to train a classifier on bag of words features to predict the topic of a specific article. Then, one can use the weights outputted by the classifier for each word as the topical score of words. For instance, if the word 'painter' appears only in articles labeled as belonging to the topic 'Arts', one might expect the classifier to assign a high score to the word painter, since it is a useful discriminating feature. As such it would be a good candidate as a topical word.

5- LDA: The LDA model (Latent Dirichlet Allocation) is completely different from all the ones considered so far. It is an unsupervised algorithm that infers its own topics, and aassigns scores for each word in each topic. Therefore, it would be interesting to compare how our hand-labeled topics compare to LDA ones.

1.3 External task: Topic classification on the Reuters dataset

To evaluate each technique, we shall use an extrinsic task. The Reuters dataset is arguably the most used dataset for topic classification. It is comprised of 21000 articles, already divided into training and testing sets, with 90 different topics, each article belonging to at least one topic, possibly more.

We will use the techniques mentioned above, i.e. log odds / signature / regression, bag of word weights, and LDA weight for 100, 200, and 300 topics or dimensions. These topics correspond to descriptors of the New-York Times. A list of these descriptors to understand to which descriptor each dimensions correspond can be found in the data folder. These descriptors were ordered by frequency of appearance in the New-York Times. We ensured that these topics were not redundant by merging topics which shared more than 50 % of their articles.

The features we feed our classifier (Linear SVC) are the mean representation of a text, i.e. we average the representation of each word in the article. As a first baseline we included a bag of words model. Also we included more widely-used distributed representation of words embeddings.

Model	**Micro F1**	**MacroF1**	**Recall micro**	**Precision micro**
BOW	0,862	0,475	0,82	0,911
Glove 100	0,799	0,341	0,735	0,875
Glove 200	0,826	0,399	0,779	0,879
Glove 300	0,837	0,429	0,793	0,886
log odds descriptors 100	0,798	0,319	0,734	0,874
log odds descriptors 200	0,82	0,397	0,778	0,867
log odds descriptors 300	0,824	0,421	0,792	0,858
Word2Vec 300	0,84	0,397	0,78	0,901
BOW from NYT 100	0,677	0,239	0,586	0,803

BOW from NYT 200	0,727	0,302	0,642	0,8375
BOW from NYT 300	0,775	0,343	0,698	0,872
LDA 100	0,41	0,11	0,54	0,33
LDA 200	0,462	0,18	0,602	0,375
LDA 300	0,513	0,27	0,657	0,421
Signature 100	0,39	0,154	0,572	0,302
Signature 200	0,49	0,201	0,705	0,385
Signature 300	0,532	0,235	0,736	0,417
Regression 100	0,497	0,177	0,613	0,421
Regression 200	0,608	0,22	0,696	0,545
Regression 300	0,66	0,241	0,572	0,781

One can see that the log odds techniques performs the best, and gives comparative results to widely used word embeddings such as Glove or Word2Vec. We conclude that the log odds technique is the best way to assign topic scores to words.

1.4 Extending the lexicon

Until now we have not addressed the issue of coverage. Our topical representation of words represents each word as a d-dimensional vector, with each dimension representing a score reflecting the association of a word to a topic. Nevertheless, our lexicon suffers from 2 coverage problems. It does not include enough words, and for the words it includes, we do not have scores for all dimensions. Although the New-York Times is big, one notices that it is by no means exhaustive. The Glove representation offers representations for 400 000 words, and Word2Vec for 3 million words. Our lexicon offers only representations for 60 000 words. What is more, these representations are often incomplete. If in a specific topic, i.e. for a specific dimension, a word has not appeared a sufficient number of times, or has not appeared at all, there is no topical score for this word in this particular topic. On average each word has only 40 % of its dimensions covered. When we did not have a score for a dimension of a word, we conveniently addressed this problem by setting its value to 0. Nevertheless, one can see the obvious utility of obtaining more fine grained score even for these dimensions. This would allow our lexicon both to extend to more words, and to be more complete.

In order to achieve this extension, we took as features glove embeddings, as target the log odds score of our seed words, and trained a regularized regression model (Ridge). The average r-square obtained for our topics is .84. This result is encouraging, meaning that it is possible to make a more complete version of our topical lexicon using this framework.

Glove embedding has been found to work better than Word2Vec embeddings for this task, although by a little margin.

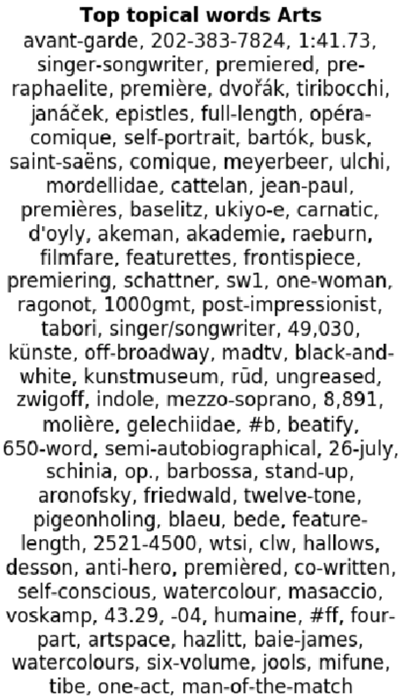
To prove the usefulness of this framework, we re-applied our topical representation of words to the text classification task, and obtained encouraging results that can be found bellow.

Model	**Micro F1**	**MacroF1**	**Recall micro**	**Precision micro**
BOW	0,862	0,475	0,82	0,911
Glove 100	0,799	0,341	0,735	0,875
Glove 200	0,826	0,399	0,779	0,879
Glove 300	0,837	0,429	0,793	0,886

log odds descriptors 100	0,798	0,319	0,734	0,874
log odds descriptors 200	0,82	0,397	0,778	0,867
log odds descriptors 300	0,824	0,421	0,792	0,858
log odds descriptors 100 augmented	0,797	0,3714	0,746	0,856
log odds descriptors 200 augmented	0,825	0,399	0,776	0,881
log odds descriptors 300 augmented	0,839	0,442	0,805	0,876
Word2Vec 300	0,84	0,397	0,78	0,901

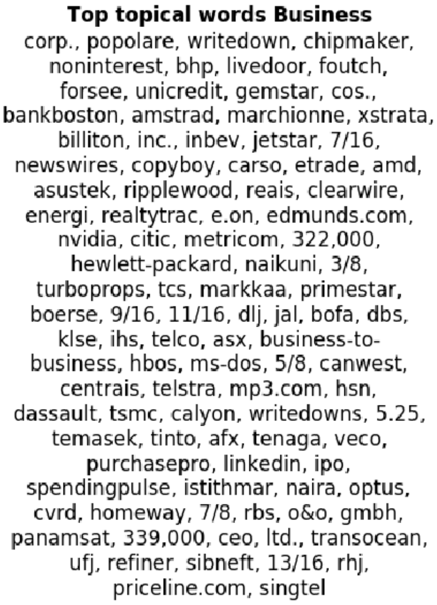
1.5 Example of topical words that are out of vocabulary

Once we trained a model for each topic, it is possible to use this model to predict which are the most topical words. Taking all the words indexed in Glove one can obtain scores for words that are out of vocabulary, but for which we have a word embedding. Thus, one can sample for each topic the top-k words as outputted by the model. Below are some results:



Top topical words Arts

avant-garde, 202-383-7824, 1:41.73, singer-songwriter, premiered, pre-raphaelite, première, dvořák, tiribocchi, janáček, epistles, full-length, opéra-comique, self-portrait, bartók, busk, saint-saëns, comique, meyerbeer, ulchi, mordellidae, cattelan, jean-paul, premières, baselitz, ukiyo-e, carnatic, d'oyly, akeman, akademie, raeburn, filmfare, featurettes, frontispiece, premiering, schattner, sw1, one-woman, ragonot, 1000gmt, post-impressionist, tabori, singer/songwriter, 49,030, künste, off-broadway, madtv, black-and-white, kunstmuseum, rūd, ungreased, zwigoff, indole, mezzo-soprano, 8,891, molière, gelechiidae, #b, beatify, 650-word, semi-autobiographical, 26-july, schinia, op., barbossa, stand-up, aronofsky, friedwald, twelve-tone, pigeonholing, blaeu, bede, feature-length, 2521-4500, wtsi, clw, hallows, desson, anti-hero, premièred, co-written, self-conscious, watercolour, masaccio, voskamp, 43.29, -04, humaine, #ff, four-part, artspace, hazlitt, baie-james, watercolours, six-volume, jools, mifune, tibe, one-act, man-of-the-match



Top topical words Business

corp., popolare, writedown, chipmaker, noninterest, bhp, livedoor, foutch, forsee, unicredit, gemstar, cos., bankboston, amstrad, marchionne, xstrata, billiton, inc., inbev, jetstar, 7/16, newswires, copyboy, carso, etrade, amd, asustek, ripplewood, reais, clearwire, energi, realtytrac, e.on, edmunds.com, nvidia, citic, metricom, 322,000, hewlett-packard, naikuni, 3/8, turboprops, tcs, markkaa, primestar, boerse, 9/16, 11/16, dlj, jal, bofa, dbs, klse, ihs, telco, asx, business-to-business, hbos, ms-dos, 5/8, canwest, centrais, telstra, mp3.com, hsn, dassault, tsmc, calyon, writedowns, 5.25, temasek, tinto, afx, tenaga, veco, purchasepro, linkedin, ipo, spendingpulse, istithmar, naira, optus, cvrd, homeway, 7/8, rbs, o&o, gmbh, panamsat, 339,000, ceo, ltd., transocean, ufj, refiner, sibneft, 13/16, rhj, priceline.com, singtel

We notice encouraging results. This sample of words with highest scores is consistent with what we would expect. We obtain words that are heavily linked with a specific topic, such as genre movements for Arts, or company names for Business. This result echoes the general behavior of the log odds scoring method. Only words that are specifically linked with a topic appear in this list, instead of common words that are more used in a specific context.

1.7 Conclusion

We built a topical lexicon that maps each word to a 100, 200, 300 vector, representing the word’s strength of association to 100, 200, and 300 selected topics. The extended version of this topical lexicon – using glove features for missing topics or dimensions - with 100, 200, and 300 dimensions for roughly 60 000 words is present in the Data folder.

2. Applications of the topical lexicon for words: selecting core words of a language

2.1 Core words of a language

Core words of a language are defined as the words used to define pervading and universal concepts. Such words include ‘school’, ‘table’, and the like. They do not convey any information other than these universal concenpts, such as information about the style of the text, or the topic discussed. They are especially useful for language learners in that they constitute the very skeleton of a language. These words are the building blocks for representing meaning in a language. Therefore, identifying such core words on which to focus study has been a task with major stakes.

2.2 Extracting core words using topical representations

The usual way to identify these core words, is to take a corpus, not unlike our corpus of the New-York Times, compute raw frequencies for each word, and then select a subset of the x most common words than constitute say 80% of all words used in the corpus. The disadvantage of such a technique is that the corpus used usually is heavily biased towards a few specific topics - the New-York Times is biased towards World Politics for instance. This technique will yield not the core words of English, but more core words of English politics and international affairs. Any measure based solely on raw frequency misses the point.

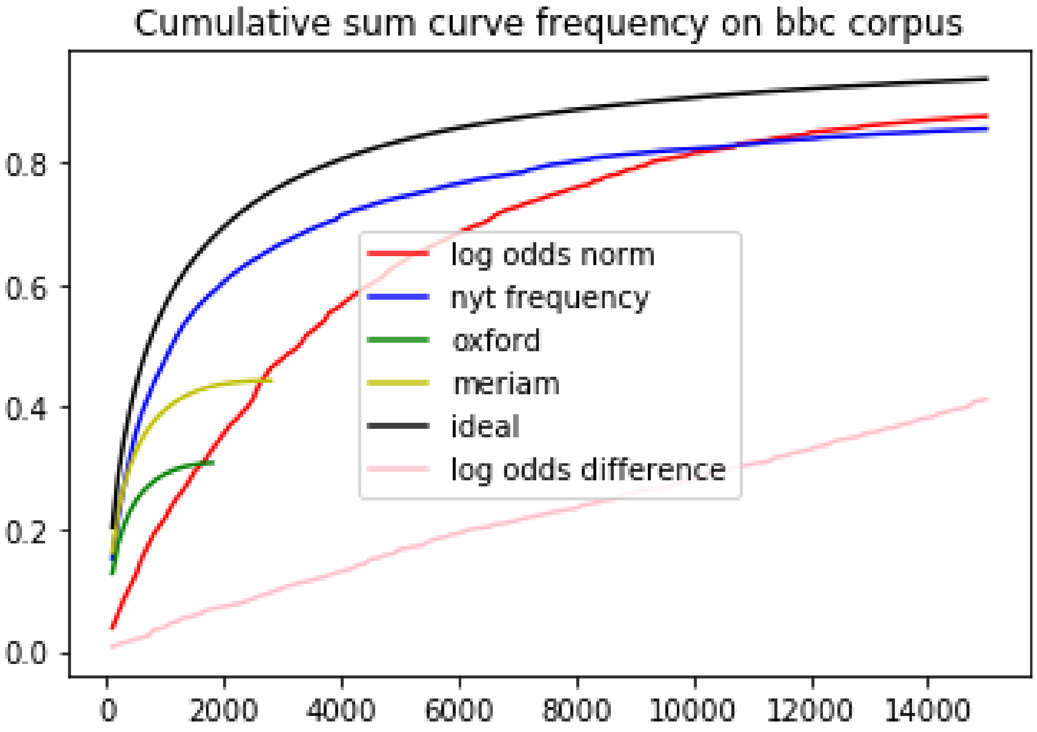
Here the approach we propose is different, and distances itself from frequency. It relies on the idea of topicality of words. True core words of English appear in roughly the same frequency no matter the coprus, and as such should not be heavily linked to any topic. To exploit this intuition, we will use our topical representation of words. So far, our technique projects each word in a d-dimensional space where each dimension is a specific topic. Building on this idea, core words of English should be vectors very close to the origin. Indeed, since they share the same distribution no matter the corpus, for each dimension they should have relatively low log odd scores. They are not linked to any specific topic.

Here we will use the extended version of our lexicon, since it allows to not deal with the problem of words that have more represented dimensions than others.

2.3 Evaluating Core words: cumulative frequency on out of sample corpus

A problem arises as we try to evaluate these models. How to say which list of core words is better than another? For this, we will rely again on the idea of frequency, although as we have stated before it is not ideal. The evaluation method will be to compute cumulative frequencies of the selected core words in a new corpora. How much of the corpus is captured by these lists of core words? This evaluation technique is tantamount to asking, if I know this list of vocabulary, how comfortable will I be in this new setting? How many words will I recognize and be familiar with in this new setting? As an out of sample corpus, we will use a large BBC dataset, with precomputed frequencies for each word.

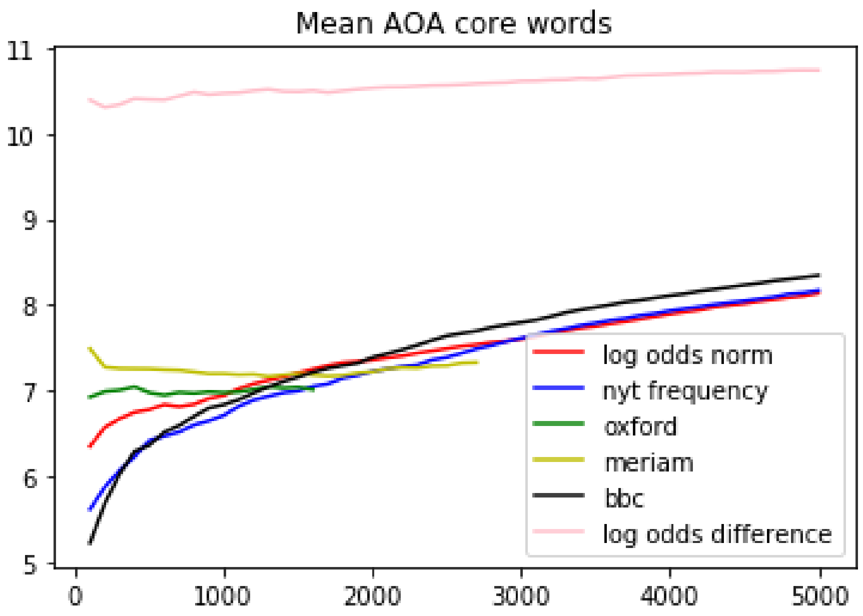
We selected 3 different baselines. The first is the core vocabulary as selected from the most frequent words in the New-York Times. The second and third are manually created lists of core words, one from the Oxford English Dictionary, the other from Meriam-Webster. We added a third technique based on our log odds score, computing the difference between the highest value in the vector and the lowest. However this technique does not perform as well.



We can see that the log odds technique performs very well, better than the 3 baselines. It is very effective in capturing words with high frequencies, and does not overfit to a particular corpus. Arguably it is better than the technique based usually on frequency. Nevertheless, one can argue that this setting is not fair either for the Meriam Webster or the Oxford core words. Indeed, we selected our set of core words based on a news corpus, and we test it on another news corpus. The Meriam Webster and Oxford lexicons are aimed at general English, English used in all contexts: TV, books, daily life, and arguably news is a small subset of all English. We leave this are to further investigation by lack of time.

2.4 Evaluating Core words: mean Age of Acquisition

Age of Acquisition is an important variable in psycholinguistic, referring to the age at which a word is typically learned. Studies in psycholinguistics show that correlates well with the speedof reading words or of naming pictures. Furthemore words that are more frequent, shorter, more familiar and refer to concrete concepts are learned earlier than more complex words. Therefore it is interesting to see which technique yields the lowest age of acquisiton.



We can see here that the log odds technique normalizing the vector outperforms the 2 lexicons based on newspaper data, and is competitive with hand labeled core words lexicons.

3. Useful resources

In the course of our study, we were brought to build 2 interesting datasets. Although, because of lack of time, and because the study lead us in different directions these 2 datasets were not exploited to their fullest potential.

3.1 Meriam-Webster learner definitions

The Meriam-Webster learner page provides interesting tag for words, such as ‘technical’, or ‘medical’. And furthermore, the definition of a word as found in a dictionary can be used in other ways.

Therefore to build a dictionary database, we scrawled the Meriam-Webster learner dictionary and obtained definitions for 36 000 words. The words and their definitions can be found in the Data folder.

3.2 PLOS Abstract / summary pair datasets

The PLOS – Public Library for Open Science – is a journal, mostly of medecine, that encourages its authors to publish in parallel of their articles, a small lay summary of their work. This initiative is done to make science more accessible and to encourage pluridisciplinarity. Here we will use them in a framework of text simplification. By contrasting these lay summaries with the abstract of the articl – the 2 are approximatevely of same length, and have the same purpose- it is possible to explore in an interesting way Readability.

Here we provide in the Data folder 20 000 odd abstract summary pairs, with the urls of the article.

4.0 Code and file description

4.1 Code

1 – toolbox : Necessary function to run the other codes, mainly include loading data

2 – topical_lexicon: Computes the topical representation of words

Parameters:
dimensions : number of dimensions, from 0 to 300, default 100
threshold : threshold of counts. Words in the topical corpus chose count is below this threshold are not considered, default 10

Returns:
topical representation of words: dictionary whose keys are words, and values are the vector of its topical representation

3 – topical_lexicon_augmented : Computes the topical representation with an extension based on glove features

Parameters:
words_corpus: words whose representation is needed, if none is provided then computes representation for all words present in glove, which is usually a lengthy process
dimensions: number of dimensions, from 0 to 300, default 100
threshold: threshold of counts. Words in the topical corpus chose count is below this threshold are not considered, default 10
alpha: value of the regularizer, default 1e-5
dimensions_glove: number of glove dimension representation with which to infer topical scores. To save space we only included the 100 dimensional distributed representation, default 100
return_r: Return the r_pearson and r_spearman coefficients for each topic

Return:
topical representation of words: dictionary whose keys are words, and values are the vector of its topical representation
if return_r = True: returns as well the r_pearson and r_spearman coefficients for each topic, as a list of tuple [r_pearson, r_spearman]

4.2 Datasets

1 - Extended topical representation of 50 000 words

The 300-topic representation of words. To get partial topical representation of words, e.g. for only 100 dimensions, truncate the vector
/Data/300_dim_topical.csv

[Column 1] = word
[Column 2 to end] = score of association of word to the topic

2 - PLOS Abstract summary pairs

Lay summary and abstract pair from the PLOS journal
/Data/plos_abstract_summary.csv

[Column 1] = url
[Column 2] = field, i.e. the subjournal the article was extracted from
[Column 2] = abstract
[Column 3] = summary
Number of abstract / summary pairs : 20 000

3 - Meriam webster definition

Definitions of words from the Meriam Webster Learner dictionary
/Data/meriam_webster_definitions.csv

[Column 1] = word
[Column 2] = definition in Meriam Webster learner dictionary
Number of definitions: 36 000

4.3 External datasets

descriptors_count_dimensions: folder containing counts for each topic, i.e. descriptors

aoa: age of acquisition table

BBC data: frequency count on bbc corpus

nyt_frequency: frequency count on New-York Times corpus

core_meriam: core words of English as stated by Meriam webster

core_oxford:core words of English as stated by Oxford Dictionary of English

descriptors_nyt: topic corresponding to each dimension

glove.6B.100d: glove feature to build extended representation of words