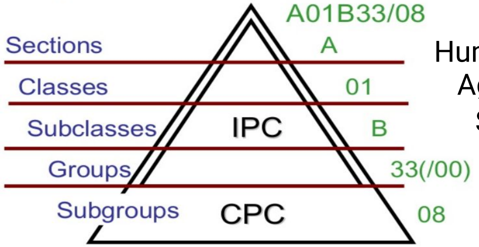
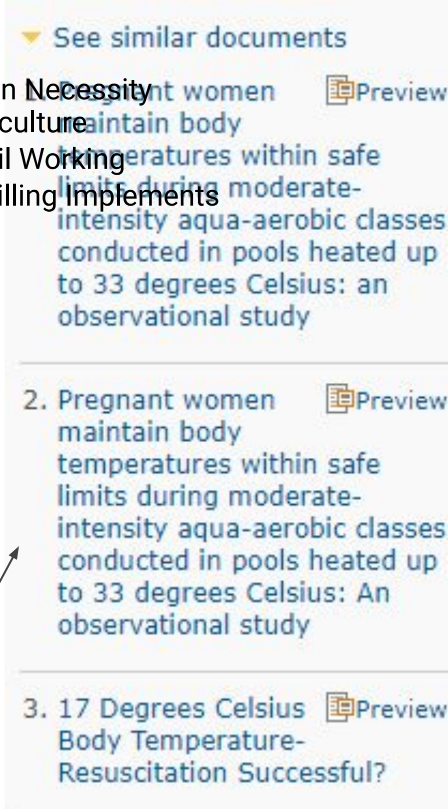
ProQuest AI Health Search

Overview & Motivation



Human Necessity
Agriculture
Soil Working
Tilling Implements

- Dialog - ProQuest Information Service
  - 1.3 Billion records
  - Powerful and efficient search for professionals
- CPC Team: Can we create a model that predicts *Cooperative Patent Classification (CPC)* tags of any given documents?
- DocSim Team: Can we enhance the user's research experience by improving the Document Similarity results?

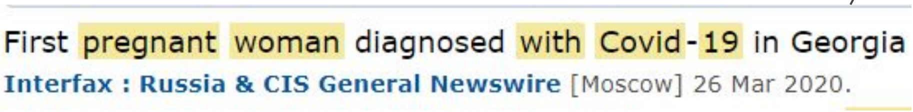


See similar documents

1. Pregnant women maintain body temperatures within safe limits during moderate-intensity aqua-aerobic classes conducted in pools heated up to 33 degrees Celsius: an observational study — Preview

2. Pregnant women maintain body temperatures within safe limits during moderate-intensity aqua-aerobic classes conducted in pools heated up to 33 degrees Celsius: An observational study — Preview

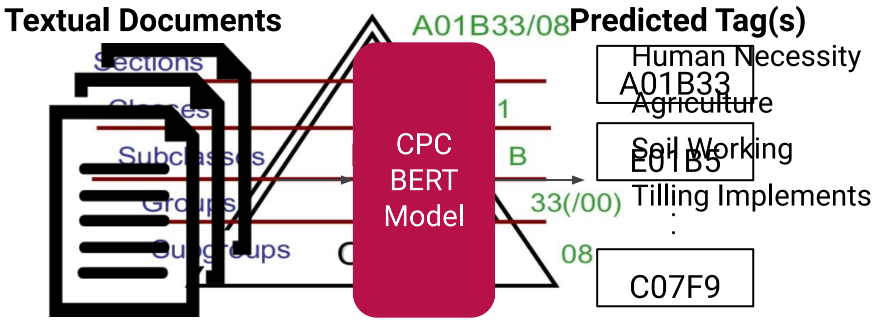
3. 17 Degrees Celsius Body Temperature-Resuscitation Successful? — Preview



First pregnant woman diagnosed with Covid-19 in Georgia

Interfax : Russia & CIS General Newswire [Moscow] 26 Mar 2020.

CPC Team



What we wanted to do:

- Create a machine learning model to predict the main group-level CPC tag of any textual documents.

What we achieved:

- Fine-tuned a BERT model to predict multiple subclass-level CPC tags of any textual documents.
  - Model achieved a normalized coverage error of 0.8.
- Created a web interface that allows users to interact with the model.

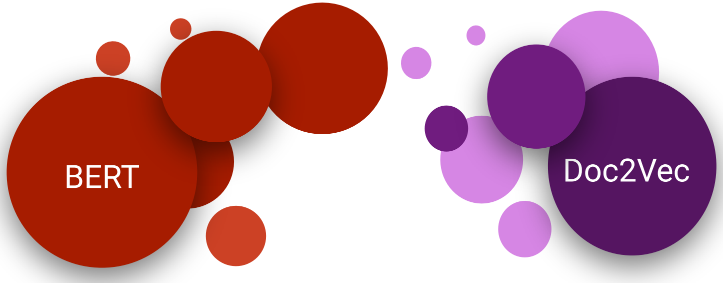
DocSim Team

What we wanted to do:

- Deploy a prototype that demos improved similarity results

What we achieved:

- Trained and fine-tuned BERT and Doc2Vec
- Clustered and evaluated
- Utilized the user’s query to rank the similarity results



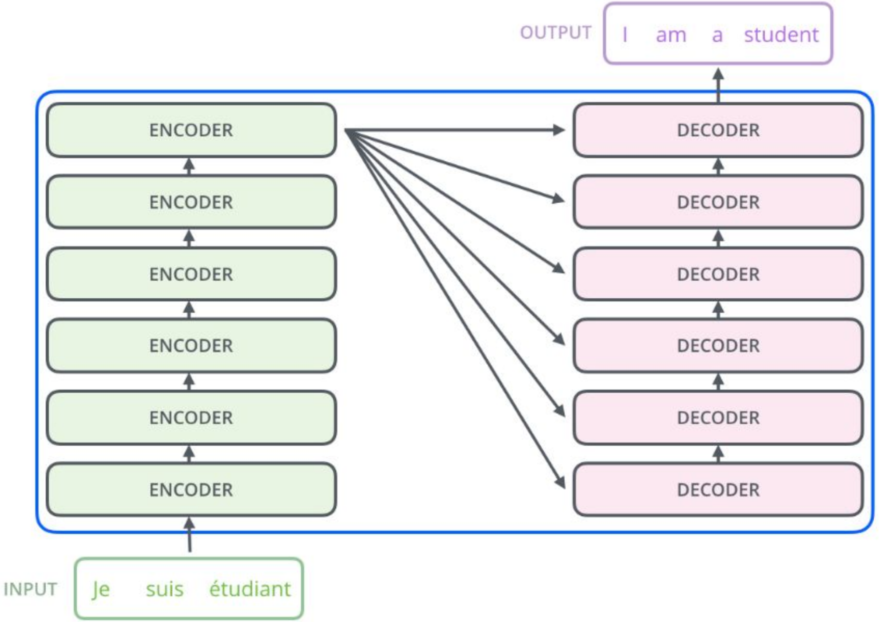
Different embeddings, different clusters

Q&A

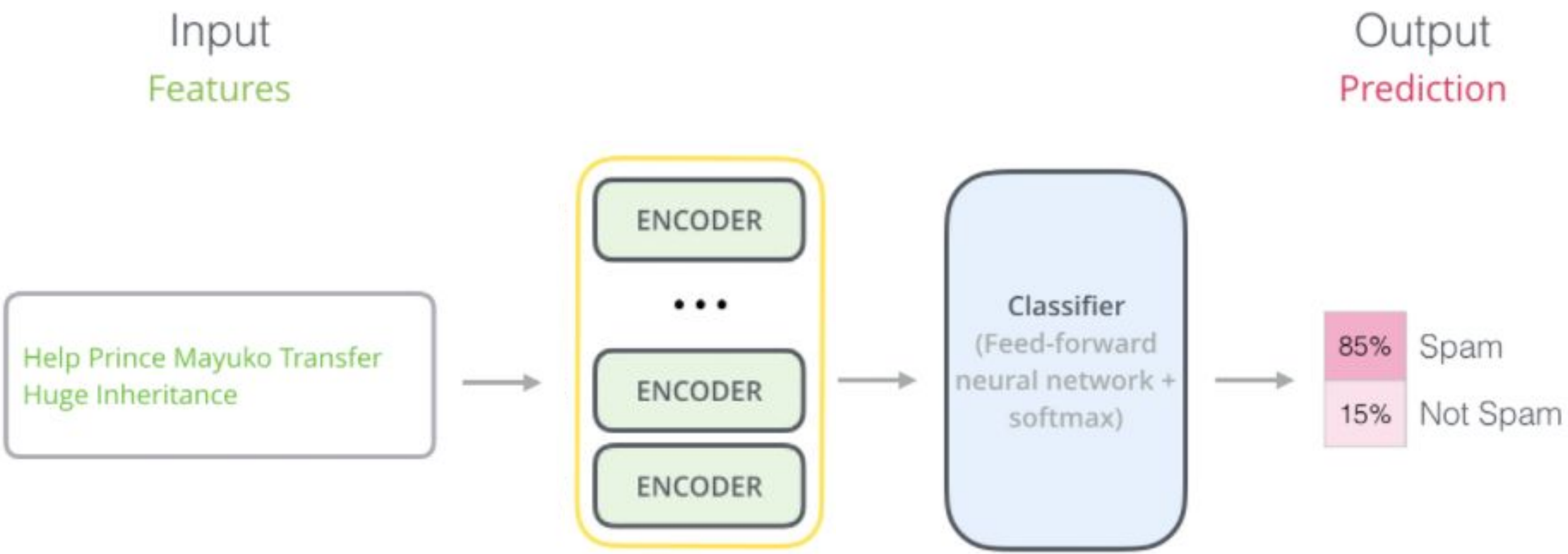
ProQuest®

Transformer

- Deep learning model based solely on attention mechanism.
- Seq2seq model (with encoder and decoder) that forgoes recurrence and convolution.



BERT


Input
Features
Help Prince Mayuko Transfer Huge Inheritance
ENCODER
• • •
ENCODER
ENCODER
Classifier
(Feed-forward neural network + softmax)
Output
Prediction
85%
Spam
15%
Not Spam

Metrics for Multi-label Classification

- **Accuracy:** percentage of documents perfectly classified. Partially correct documents are viewed as misclassified.
- **Coverage Error:** average number of labels that have to be included in final prediction (by rank) to cover all true labels. The best value for this is the average number of true labels in the test set.
- **Label ranking loss:** best think of average percentage of incorrectly ordered label pairs (i.e. a true label ranks lower than a false label). The best value is 0.
- **ROC AUC score:** think of this as how well your model does in ranking correct labels higher than incorrect labels. The best value is 1.

Coverage Error

- **Example 1:** Let the true label of a document be [0, 1, 1, 0] and let the prediction score of the model be [0.1, 0.9, 0.4, 0.8]. The coverage error for this sample would be 3 since by ranking prediction score from highest to lower, we would have to take the top 3 predictions (0.9, 0.8, 0.4) to cover all the true labels .

- **Example 2:** Let the true label of a document be [1, 1, 0, 1] and the prediction score of the model be [0.7, 0.9, 0.3, 0.8]. The coverage error for this sample is 3 because we need to take the top 3 predictions (0.9, 0.8, 0.7) to cover all true labels. With this, we can see why the best value for coverage error is the average number of true labels in your test set.

Coverage Error Illustrated

Coverage Error = # of top predictions to cover all true labels.

	True	Pred Scores
A	0	0.1
B	1	0.9
C	1	0.4
D	0	0.8

Sort by Pred →

	True	Pred Scores
B	1	0.9
D	0	0.8
C	1	0.4
A	0	0.1

True Labels = [B, C]

Top 1 Pred = [B] Missing C

Top 2 Pred = [B, D] Missing C

Top 3 Pred = [B, D, C]

Coverage Error = 3

Coverage Error Illustrated

	True	Pred Scores
A	1	0.7
B	1	0.9
C	0	0.3
D	1	0.8

Sort by Pred →

	True	Pred Scores
B	1	0.9
D	1	0.8
A	1	0.7
C	0	0.3

True Labels = [A, B, D]

Top 1 Pred = [B] Missing A, D

Top 2 Pred = [B, D] Missing A

Top 3 Pred = [B, D, A]

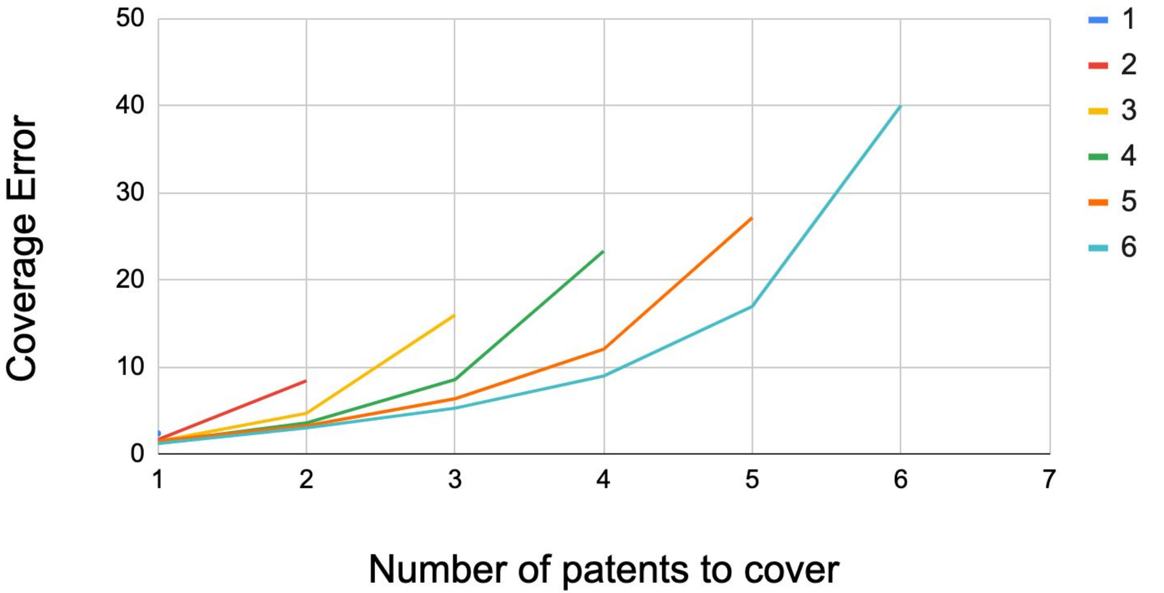
Coverage Error = 3

Issue with Coverage Error

- Coverage error reported by the model is the mean of coverage errors of the samples in val/test set.
- Different samples will have different ideal coverage error
  - If a sample has more true labels, it's ideal coverage error will be higher
- Suggestion: Normalize the coverage error

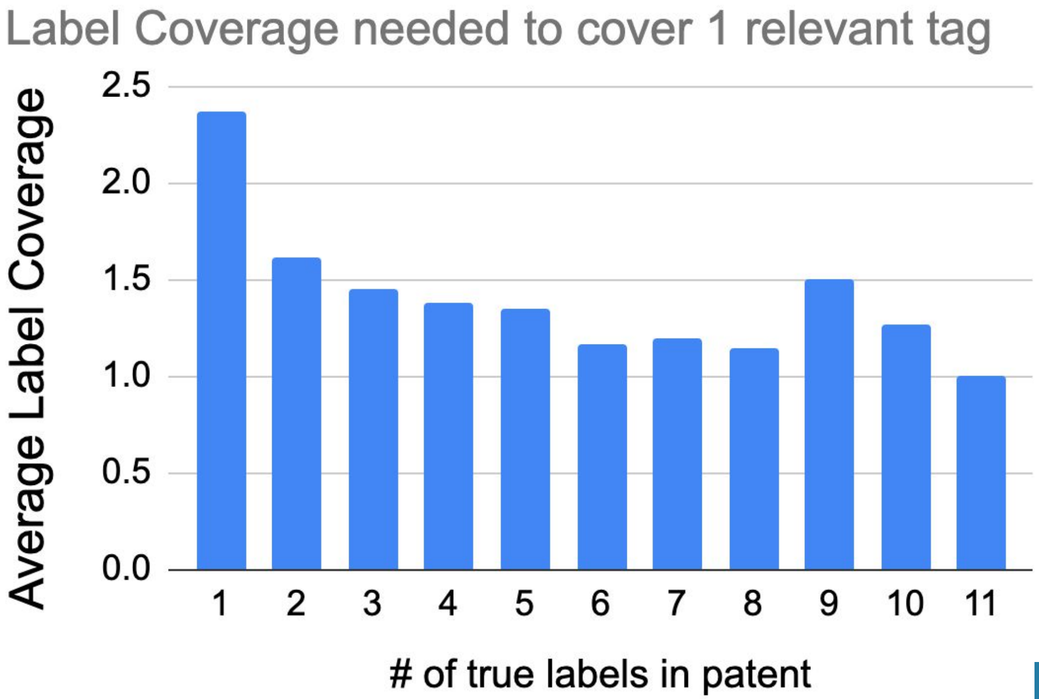
$$Normalized\ CE = \frac{\#\ of\ true\ labels}{Coverage\ Error}$$

It takes our model quite a few tries to get the last few relevant tags



Coverage error taken for increasing tag coverage for patents with x labels(legend)

For patents with a lot of labels we always get something relevant early enough

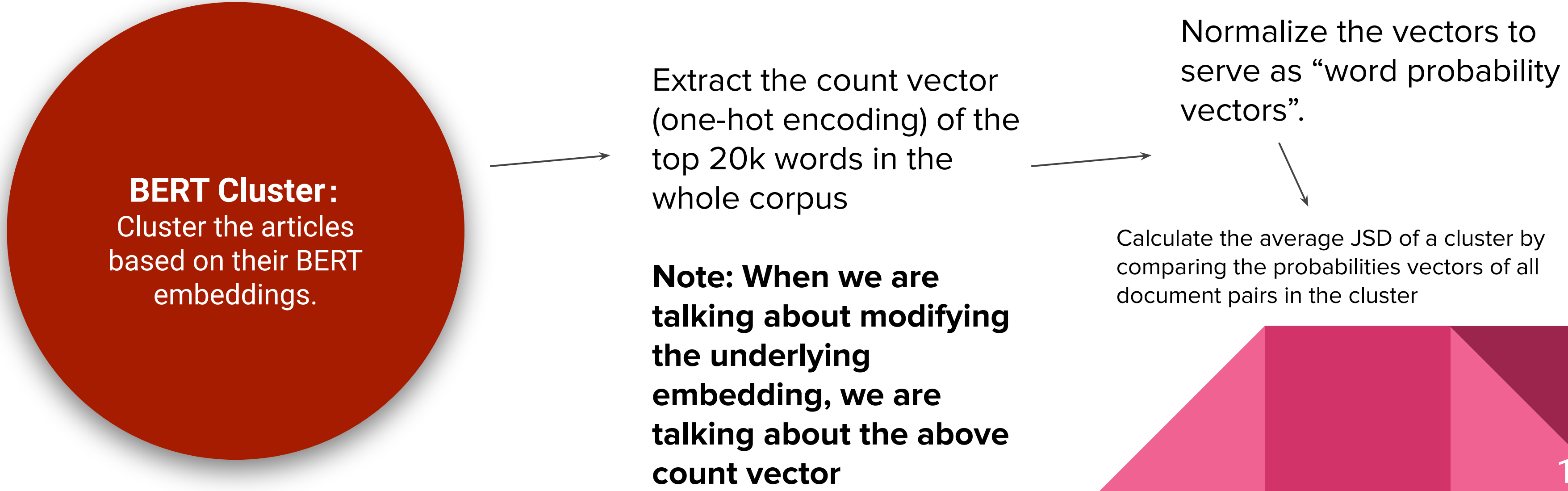


Pretty Analysis numbers

- The models top prediction is relevant 84% of times
- Model retrieves a relevant tag out of the top 3 tags it predicts 95% of times

JSD of a cluster

How similar are the documents in a single cluster? Let's use a Divergence measure to calculate. First we need to break down a document into a probability vector.



BERT Cluster:
Cluster the articles based on their BERT embeddings.

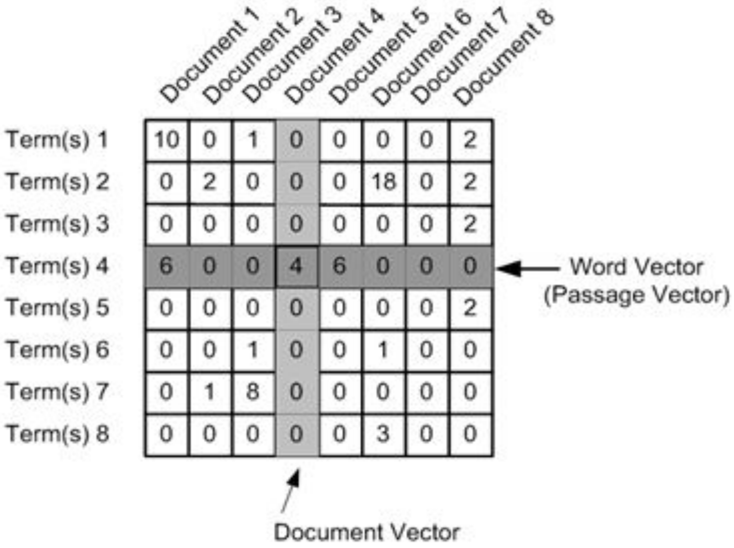
Extract the count vector (one-hot encoding) of the top 20k words in the whole corpus

Note: When we are talking about modifying the underlying embedding, we are talking about the above count vector

Normalize the vectors to serve as "word probability vectors".

Calculate the average JSD of a cluster by comparing the probabilities vectors of all document pairs in the cluster

Count Vectorizing



	Document 1	Document 2	Document 3	Document 4	Document 5	Document 6	Document 7	Document 8
Term(s) 1	10	0	1	0	0	0	0	2
Term(s) 2	0	2	0	0	0	18	0	2
Term(s) 3	0	0	0	0	0	0	0	2
Term(s) 4	6	0	0	4	6	0	0	0
Term(s) 5	0	0	0	0	0	0	0	2
Term(s) 6	0	0	1	0	0	1	0	0
Term(s) 7	0	1	8	0	0	0	0	0
Term(s) 8	0	0	0	0	0	3	0	0

Imagine these 8 documents are in a single cluster. Each cell indicates the number of the corresponding Term(s) appear in this document. Here we simply normalize each column (so they sum to 1)   and compare the term's "probability" distribution.

Instead of just 1 term represented as the row, we can use n-grams as well.

Coherence of a cluster[1]

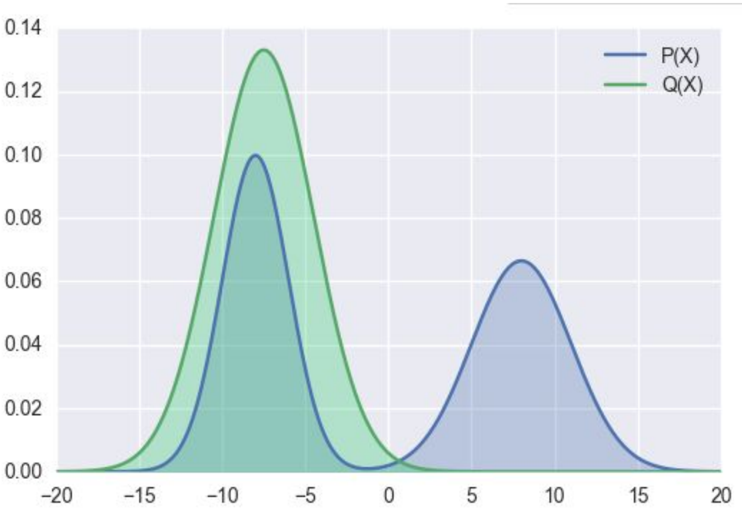
Once we have the average JSD of a cluster, we must normalize the divergence since it naturally increases with cluster size. To do this:

1. Sample a few "random clusters" within the cluster size range of the solution
2. Learn a smooth linear function for the random JSD across cluster size
    a. The r^2 value for this function is almost always > .999 with greater than 10 samples ---- the linear relationship exists and is significant.
    b. This is very helpful because now we are able to parametrically calculate the random JSD as a formula of cluster size, so that we don't need to explicitly cluster and measure for every possible cluster size that may occur, which saves us a lot of time.
3. Subtract the average JSD of the cluster from the random JSD of the same cluster size

Then you can average things together to get the final coherence score.

[1] Clustering More than Two Million Biomedical Publications: Comparing the Accuracies of Nine Text-Based Similarity Approaches

JSD



Word Prob Vector for Article P

Word Prob Vector for Article Q

K-L Divergence

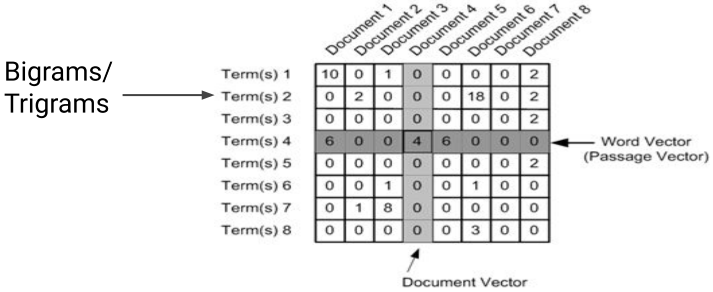
$$\mathrm{JS}(p \parallel q) = \frac{1}{2}\mathrm{KL}(p \parallel m) + \frac{1}{2}\mathrm{KL}(q \parallel m)$$

where

$$m(x) = \frac{1}{2}\left(p(x) + q(x)\right)$$

$$\mathrm{KL}(p \parallel q) = -\sum_{x} p(x) \log\left(\frac{q(x)}{p(x)}\right)$$

Bigrams/Trigrams meaning



	Document 1	Document 2	Document 3	Document 4	Document 5	Document 6	Document 7	Document 8
Term(s) 1	10	0	1	0	0	0	0	2
Term(s) 2	0	2	0	0	0	18	0	2
Term(s) 3	0	0	0	0	0	0	0	2
Term(s) 4	6	0	0	4	6	0	0	0
Term(s) 5	0	0	0	0	0	0	0	2
Term(s) 6	0	0	1	0	0	1	0	0
Term(s) 7	0	1	8	0	0	0	0	0
Term(s) 8	0	0	0	0	0	3	0	0

So, when we say we’re working with bigrams and trigrams, we’re modifying the underlying embeddings (see slide 4) that produce the word-probability vectors which are used to calculate JSD. If we use this to compare just Doc2Vec and BERT, it is unbiased because the embeddings used to create the clusters are unrelated with the way we produce the word-probability vectors.

TF-IDF, however, is a transformation of the underlying embedding (count vector), so it naturally performs better in this measure, so textual coherence is naturally biased towards TF-IDF as they use similar idea to represent each sentence/document.

Textual coherence results

Embedding | ngram	Unigram	Bigram	Trigram
BERT	2.00	1.03	0.18
Doc2Vec	2.18	**2.09**	**0.26**
TF-IDF	**3.91**	1.53	0.24

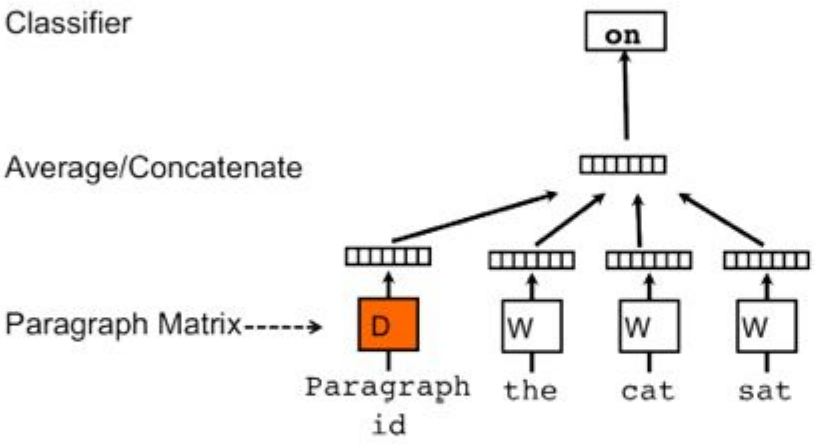
Interpretations

A 0 coherence score means that the clusters created by whatever technology are about the same unigram/bigram/trigram textually as randomly picking documents. (not good!)

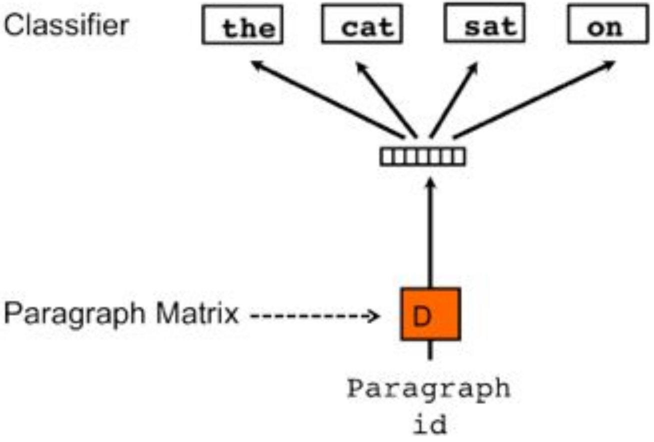
If an embedding does better relative to TF-IDF in (n>1)grams, then it has more similar longer strings of text in its clusters. If we increase the length of the underlying embeddings, then it is using more than just the top 20k ngrams in the corpus.

We want an embedding that delivers coherence to some comparable degree to TF-IDF along with our other measures.

Doc2Vec

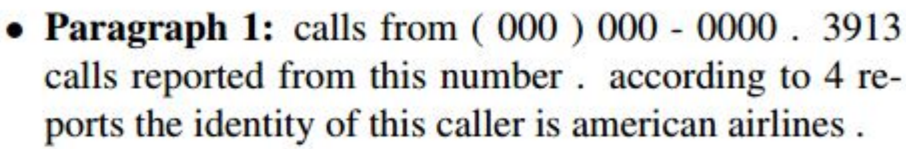


DM (Distributed Memory)

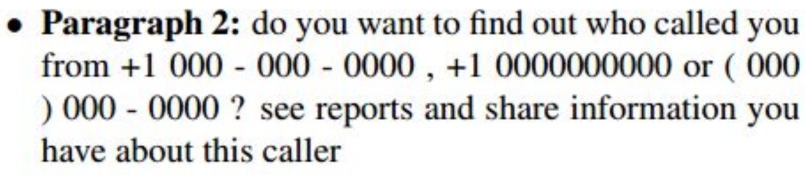


DBOW (Distributed Bag of Words)

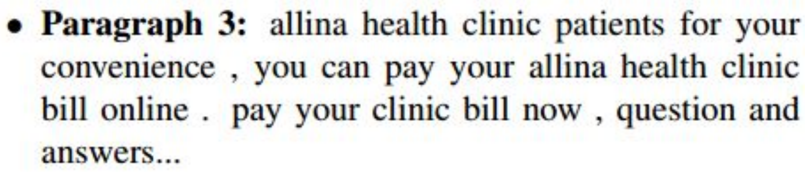
Doc2Vec Example



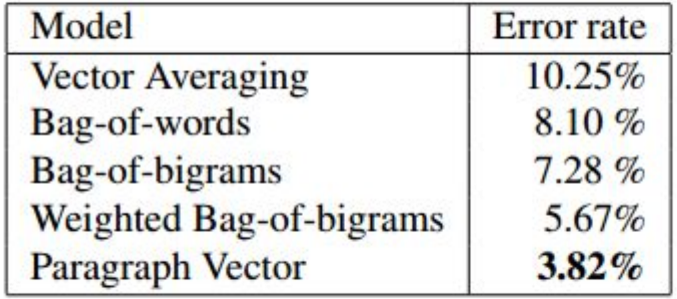
- **Paragraph 1:** calls from (000) 000 - 0000 . 3913 calls reported from this number . according to 4 reports the identity of this caller is american airlines .



- **Paragraph 2:** do you want to find out who called you from +1 000 - 000 - 0000 , +1 0000000000 or (000) 000 - 0000 ? see reports and share information you have about this caller



- **Paragraph 3:** allina health clinic patients for your convenience , you can pay your allina health clinic bill online . pay your clinic bill now , question and answers...



Model	Error rate
Vector Averaging	10.25%
Bag-of-words	8.10 %
Bag-of-bigrams	7.28 %
Weighted Bag-of-bigrams	5.67%
Paragraph Vector	**3.82%**

Query = "coronavirus vaccine"

1. These results suggest that M2e-MAP presenting M2e H5N1 virus has great potential be developed into an effective subunit vaccine prevention infection by broad spectrum HPAI H5N1 viruses

2. Middle East Respiratory Syndrome coronavirus MERS-CoV is viral respiratory disease Most people infected with MERS-CoV develop severe acute respiratory illness It was first reported Saudi Arabia 2012 has since spread several other countries We report c

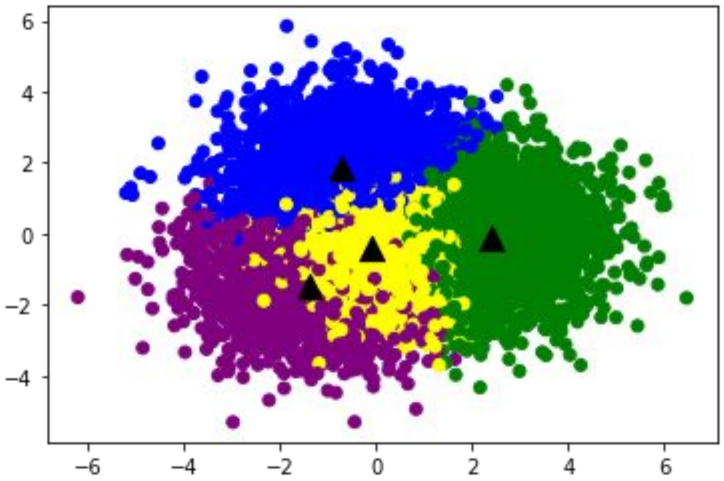
3. This study confirmed circulation influenza AH1N1 AH3N2 B viruses human population Central Africa describes emergence oseltamivir-resistant AH1N1 viruses Central Africa

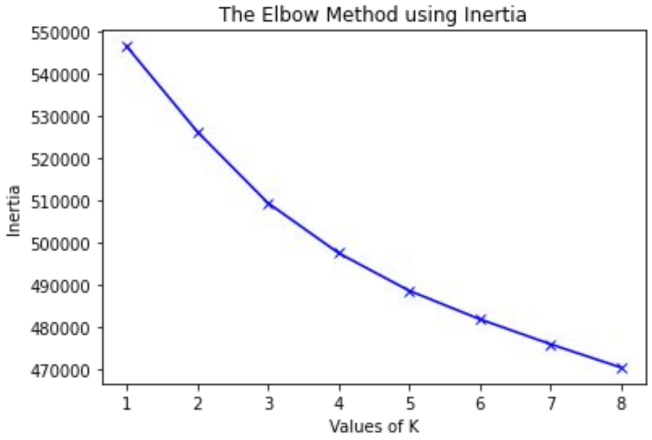
4. Vaccine Research Center US National Institute Allergy Infectious Diseases NIH

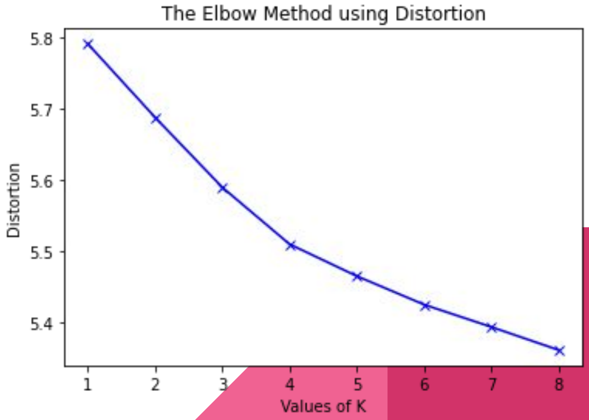
5. canary pox vector gp120 vaccine ALVAC-HIV AIDSVAX B/E gp120 RV144 HIV-1 vaccine trial conferred an estimated 31% vaccine efficacy Although vaccine Env AEA244 gp120 is antigenic unmutated common ancestor V1V2 broadly neutralizing antibody bnAbs no plasma bnAb

K-Means Clustering

Doc2Vec Clusters




The Elbow Method using Inertia


The Elbow Method using Distortion