A Dissimilarity Based Concordancer

Submitted By:
Shonna Gu
Nicholas Sanders
Komal Mistry

THE UNIVERSITY OF BRITISH COLUMBIA

About the Team

Komal Mistry	Shonna Gu	Nicholas Sanders	Prof. Miikka Silfverberg
• B.Tech in electronics and telecommunications engg from MPSTME Mumbai. • MBA in Business Intelligence and Analytics from NMIMS, Mumbai	• Bachelor in Internet of Things Engineering at Beijing Institute of Technology • Master in Computer Science at Beijing Institute of Technology	• BA in Linguistics from Western Washington University	• PhD in Language Technology from the University of Helsinki in 2016. • Postdoctoral researcher at the University of Colorado at Boulder where he worked on computational modeling of word structure using deep neural networks.

We'd also like to thank our CRIM mentors **Pierre André Ménard** and **Lise Rebout** for their constant guidance and support.

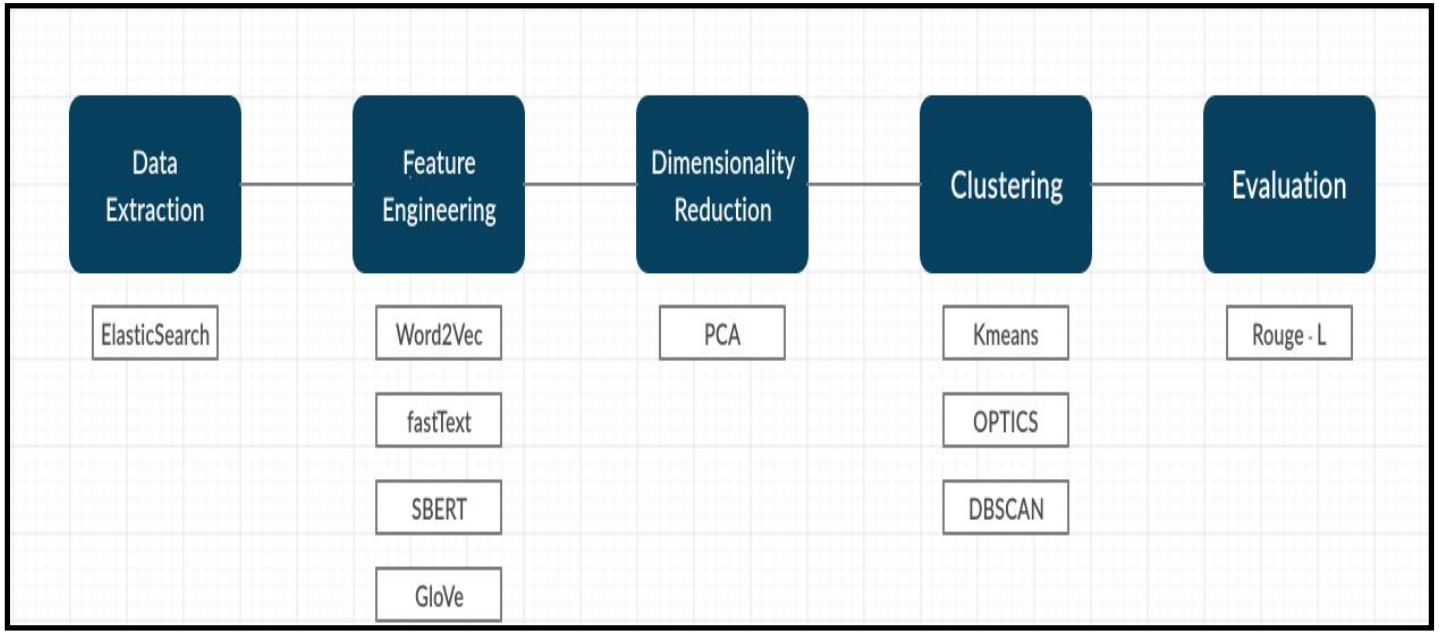
What is a Concordancer ?

A concordancer is a tool to show a target word in its context.

- A concordancer (sometimes called key word in context or KWIC) is a useful tool for linguists, computational linguists or data scientists who need to target specific query word in a text corpus to explore the cues, hints and context that underlie the use of such a phenomenon
- It aligns instances of the query word found in a corpus and shows the surrounding words, enabling the user to assess the context of the term in the corpus.
- Concordancers have traditionally been used to provide easy access to key ideas in religious scripts like the Bible, Vedas and Quran to name a few.
- With advances in technology, concordancers are now used in corpus exploration, providing linguists with a tool to study a large corpus by comparing different uses of a term, finding phrases, and examining idioms used in the text.

Current Concordancers

Project Flow



Applications of Concordancers

- Concordancers have applications in English Language Teaching (ELT) , where they are used to acquaint learners with collocations and word usage.

- Using concordances, students can learn the frequency of use and the context in which words should be used (Fatih 2014)

- Furthermore, within the field of computational linguistics, concordancers can be used to explore corpuses and enable researchers to quickly analyse a large corpus for frequency of n-grams and corpus diversity

- Many large corpora, like the American National Corpus, British National Corpus, or Corpus of Contemporary American English, are released with concordancing features enabled.

Why a dissimilarity based concordancer

- Typical and historical concordancers are similarity-based and return all matching occurrences, or concordances, for a given target word.
- This allows for a user to observe how a word appears in a corpus, but can be exhaustive, as often all concordances must be manually parsed by the user to be of any use.
- We investigate and propose an automatic alternative method for returning the most different concordances, based on dissimilarity.
- We do this by exploring different feature engineering methods for text data, unsupervised clustering algorithms, and dimensionality reduction techniques.

Phase 1 Datasets

datasets	name	language	document count	token count	Annotated or not
Travaux de l’assemblée nationale	Assnat	French	5,484	231M	yes
COVID-19 Open Research Dataset	Covid	English	57,368	295M	yes

Note: the datasets are quite big that each one takes ~80G space after decompression.

Phase 1 Datasets

```
- corpus.json
- CorpusStructure.json
- documents
        - 0a01ca28-7ebc-11ea-b2fa-02420a0000bf.json
        - 0a1e631c-7ebc-11ea-9100-02420a0000bf.json
        - ...
 - groups
        - 757fc491-0f25-4ad7-914c-cd906afa7f13
            - 0a01ca28-7ebc-11ea-b2fa-02420a0000bf.json
            - 0a1e631c-7ebc-11ea-9100-02420a0000bf.json
            - ...
        - 2b830e5f-f607-4ff8-b38e-3a4d028cced8
            - 0a01ca28-7ebc-11ea-b2fa-02420a0000bf.json
            - 0a1e631c-7ebc-11ea-9100-02420a0000bf.json
            - ...
```

corpus.json

- basic information of this corpus
- e.g. corpus id, corpus title, language, document counts

CorpusStructure.json

- detailed structure of ‘groups’

documents

- file names are document ids
- each document includes an id, source, title of the document, *full text* and the language

groups

- annotation files
- one group is about document metadata
- another is about split sentences and tokens

Phase 1 Data Retrieval using ElasticSearch

- **What kind of data we need**
  - query word and its contexts
  - *e.g. oral administration only elicited an IMMUNE response in one of six*
- **What difficulties we have when retrieving data**
  - no enough space for decompression
  - a linear search function takes an incredibly long time to execute
- **How we solved the problems**
  - read data from zip file(~7G) directly
  - built Elasticsearch indices for quickly searching targets
- **How we used Elasticsearch in this project**
  - created indices upon each document and retrieved all documents that contain a query word
  - extract instances by iterating these retrieved documents

Note: Elasticsearch is an open source distributed, RESTful search and analytics engine capable of efficiently solving search and querying for large datasets

Phase 2 Feature Engineering

- This phase of the pipeline consists of deriving meaningful features from the contexts / instances retrieved from the elasticsearch querying
- The goal is to vectorize this collection of instances so that we can use clustering to return a diverse set of concordances, based on the features represented in the vectors.
- In our implementation we consider each context to be an aggregated vector which is formed from the vectors of the words found in the context. To explore the efficiency of word embeddings, we chose four techniques to extract vectors:

    a. Word2Vec (model derived from 10,000 documents in COVID and 1000 documents in Assnat corpus)

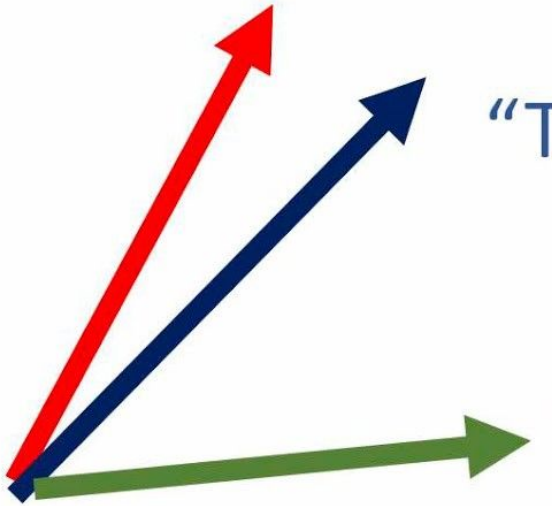
    b. GloVe (Use pre-trained embeddings to overcome the issue of local statistics faced in the Word2Vec model)

    c. fastText (Use pre-trained embeddings to overcome OOV problem by utilizing character level embeddings)

    d. SBERT (Use pre-trained embeddings , current state of the art technique for extracting semantically meaningful vectors)

What Is a Word Vector

"Lion is the king of the jungle."

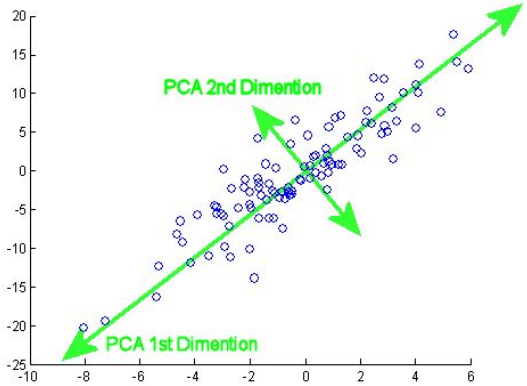


"The tiger hunts in this forest."

"Everybody loves New York."

Phase 3 Dimensionality Reduction

Principal Component Analysis (PCA)



https://medium.com/@raghavan99o/principal-component-analysis-pca-explained-and-implemented-eeab7cb73b72

- It finds the directions of maximum variance in the higher dimension and projects this variance into a lower dimension , thus preserving most of the observed variance from the original word embeddings, while reducing the number of dimensions of the word embeddings

Phase 4 Clustering

What is clustering in unsupervised learning

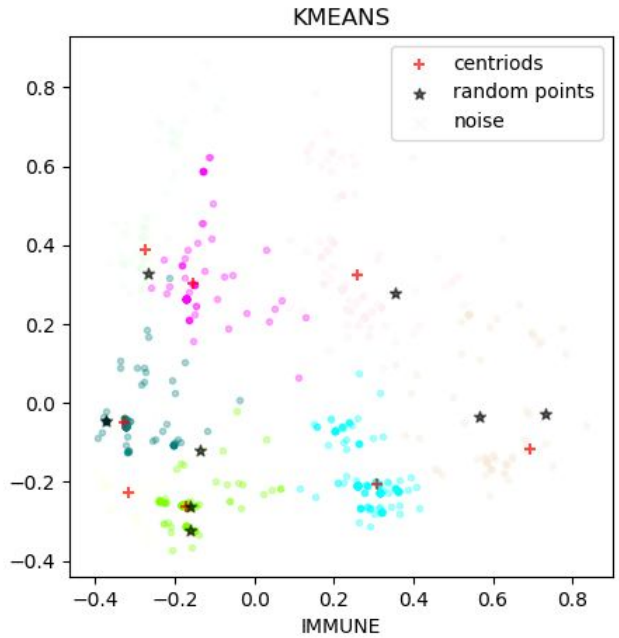
- An algorithm that can group or cluster data points based on position in space using a given distance metric (we use cosine similarity)

Why do we choose clustering

- Clustering is used so that we can classify data points and maximize diversity of concordances by returning each cluster's most center point (the centroid)

Different clustering methods examined

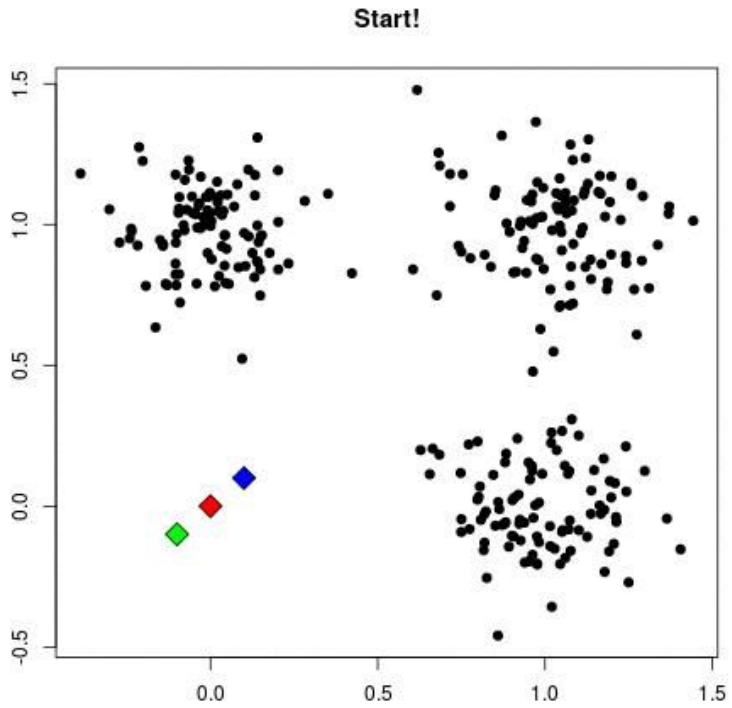
- We examined KMeans, DBSCAN, and OPTICS from scikit-learn, which all come with different performance implications (will be explained further)



Phase 4 Clustering - KMeans

How does KMeans work

- Randomly initiate centroids and assign data points to their closest centroid, recalculate the cluster centroids based on the mean of all data points within each cluster and repeat



https://i.imgur.com/k4XcapI.gif

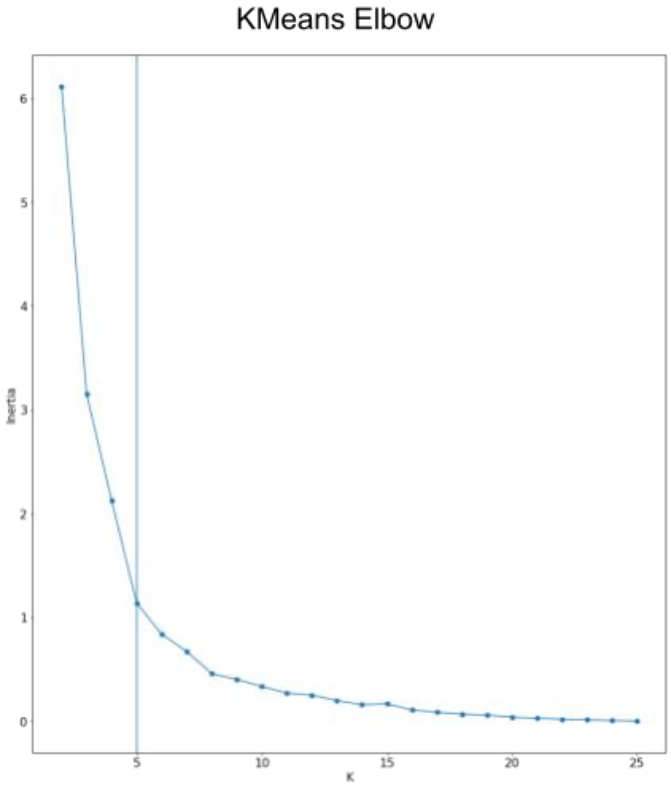
Phase 4 Clustering - KMeans

How are we choosing the hyperparameters

- 'K' clusters determined by automatically finding the elbow in a distribution of inertias (the avg. distance of data points from centroids)
- Iterations set to 300 (default)

Considerations:

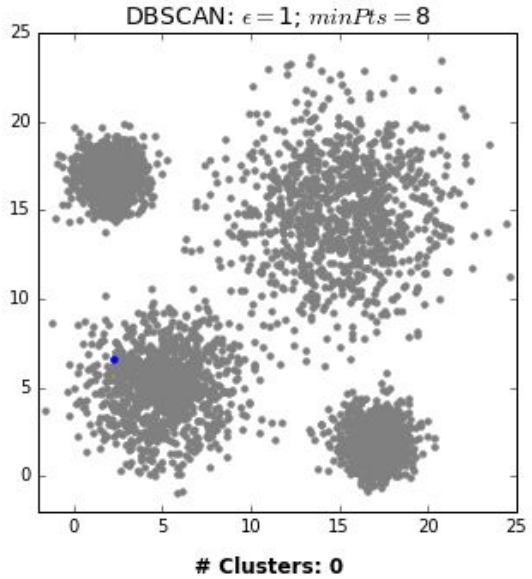
- Some randomness and sensitive to noise

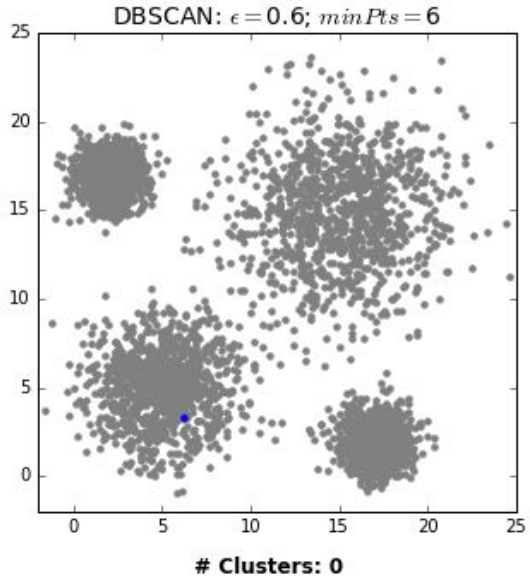


Phase 4 Clustering - DBSCAN

How does DBSCAN work

- Density Based Spatial Clustering of Applications with Noise. Using its hyperparameters and given metric to determine clustering.





https://dashee87.github.io/images/DBSCAN_search.gif

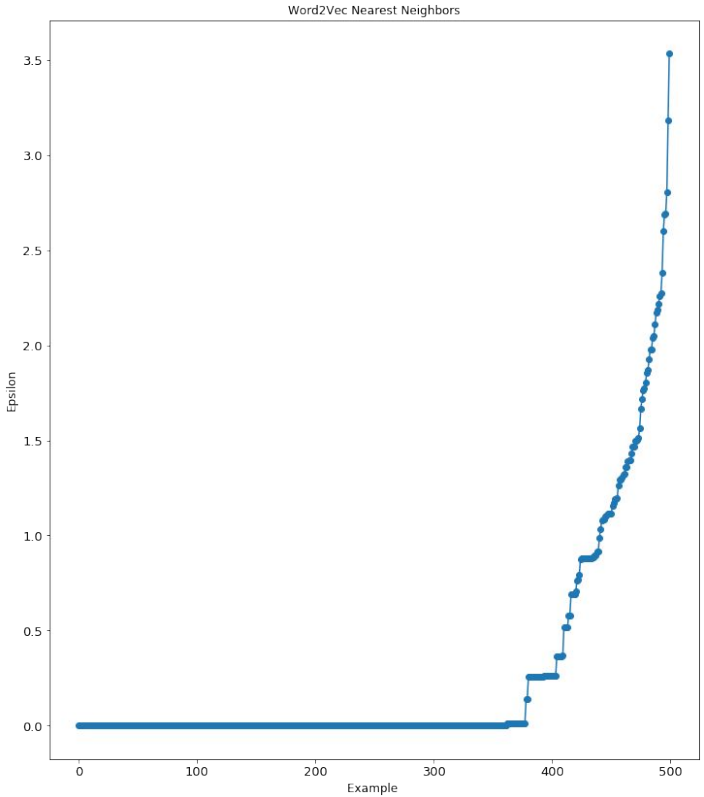
Phase 4 Clustering - DBSCAN

How are we choosing hyperparameters

- Determining epsilon (measured in euclidean distance) and minimum samples with distance of nearest neighbors and trial and error

Considerations:

- Epsilon and minimum samples are hard to determine and sensitive to other system parameters



Phase 4 Clustering - OPTICS

How does OPTICS work

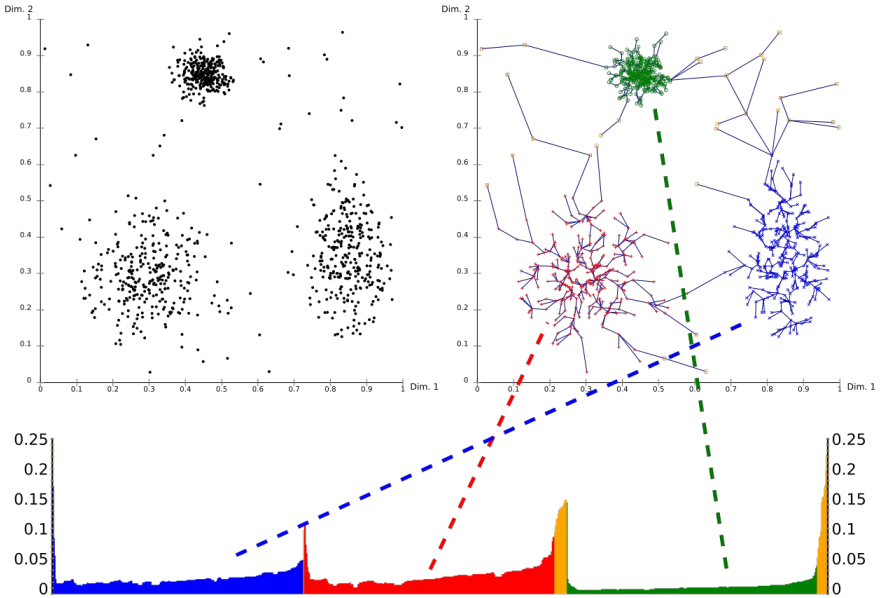
- Ordering Points to Identify Cluster Structure. Value is set for min samples, calculates a core distance and reachability distance for each data point, and then updating cluster centroids

How are we choosing hyperparameters

- Minimum samples chosen through trial and error based on other parameters

Considerations:

- Less hyperparameters can mean less control



By Chire - Own work, Public Domain, https://commons.wikimedia.org/w/index.php?curid=10293701

Evaluation

- After clustering the word embeddings , we extract the cluster centers or take random samples from each cluster as per clustering technique. This new set of samples is then treated as our final context diverse concordance result for a given query word.
- We use a metric called **ROUGE (Recall Oriented Understudy for Gisting Evaluation)** to evaluate the context diversity of the resulting concordances.
- Rouge metric is calculated by counting the ngram overlaps occurred between two text samples.

```
from rouge import Rouge

hypothesis_a = "The virus can cause pneumonia"
hypothesis_b = "The virus has corrupted the disk memory"
reference= "The virus can cause pneumonia"

rouge = Rouge()
scores_a = rouge.get_scores(hypothesis_a, reference)[0]['rouge-l']['f']
scores_b = rouge.get_scores(hypothesis_b, reference)[0]['rouge-l']['f']

print(scores_a, scores_b)
```

```
0.999999995 0.33333332847222225
```

Evaluation

- We would expect the ROUGE score before clustering to be higher than the ROUGE score produced after clustering and in general for any given concordance output we would like to observe lower ROUGE scores to indicate that the samples are context diverse.

- We use the ROUGE - LCS variant, LCS standing for longest common substring. The reasoning behind it being that if two text samples were to have the same phrase or LCS then it’s most likely that they have used the query word in the same context.

Results

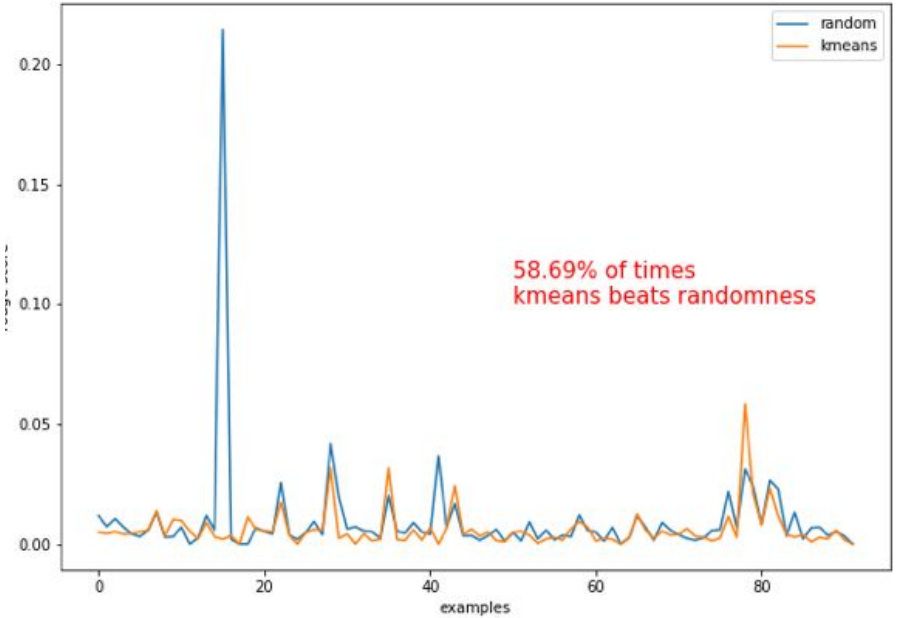
Query words	Number of clusters	Cluster results	Rouge score-KMeans	Random results	Rouge score-Random
PAUCITY	16	normal . The **PAUCITY** of lung findings was . There is a **PAUCITY** of bat-borne virus break the of **PAUCITY** severe of acute new reflect risk **PAUCITY** of this studies end sidering the **PAUCITY** of information avai ion The type **PAUCITY** . Cross-provincial dge gaps and **PAUCITY** of scientific data o . A relative **PAUCITY** of resident alveolar oglia . This **PAUCITY** of Ia Ag on astro lly the by , **PAUCITY** of data to level v , an overall **PAUCITY** of basic residues rently mice **PAUCITY** studies lack microb 44 subjects **PAUCITY** in CD4+ studies ( ) s where data **PAUCITY** is present and impli 7] There is **PAUCITY** of population-based mediated on **PAUCITY** the of virus emb	0.2375	their relative **PAUCITY** emphasizes the clinical use , **PAUCITY** of evidence-bas rveillance and **PAUCITY** of disease burd Further , the **PAUCITY** of asthma hospi t . There is a **PAUCITY** of data surroun ons there is a **PAUCITY** of data about t . . 2014 to US **PAUCITY** estimated data ed , given the **PAUCITY** of laboratory-b y , there is a **PAUCITY** of comparative le where snail **PAUCITY** present and imp ly reflect the **PAUCITY** of clinical stu 20] there is **PAUCITY** of data on the systems , the **PAUCITY** of animal-infec o to influenza **PAUCITY** (reliable et c B1 vaccine is **PAUCITY** chemically LBP ozoa , and the **PAUCITY** of ciliates , r	0.2875

Note: KMeans + Sbert + PCA dim = 16 + window size = 1

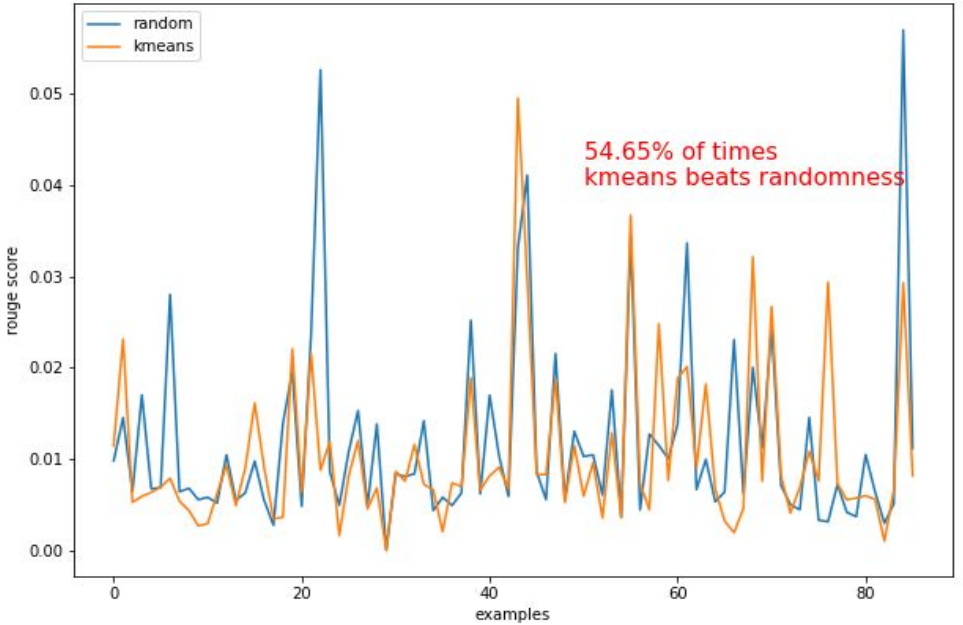
Results

Query words	Number of clusters	Cluster results	Rouge score-KMeans	Random results	Rouge score-Random
INFECTION	25	earch on viral **INFECTION** indicated that a on , including **INFECTION** of humans . HPAI ion acute hMPV **INFECTION** . During the tim s of childhood **INFECTION** appear to be mil ve evidence of **INFECTION** . Immunofluoresc RS coronavirus **INFECTION** (Lee et al. 200 ource of human **INFECTION** with Middle East read safety it **INFECTION** be activities wi supported MHV **INFECTION** of transfected 2 and rotavirus **INFECTION** ; IgG ASCs were ous studies as **INFECTION** and disease mode information on **INFECTION** control , we als and abdominal **INFECTION** will cause pain luenza A virus **INFECTION** in cats in China are-associated **INFECTION** and the incidenc) . During an **INFECTION** , the host prote owing neonatal **INFECTION** . This , in turn C3 without DV2 **INFECTION** (Fig. 1A-b) . The day before **INFECTION** , RAW264.7 cells	0.0267	related to MERS **INFECTION** as it is kno suggested PEDV **INFECTION** could be mor e prevalence of **INFECTION** varies betw were the after **INFECTION** microinjecti persistence of **INFECTION** is ability o pected be virus **INFECTION** Before Howev s simplex virus **INFECTION** of Vero cell diated p.i. was **INFECTION** in with comp ute respiratory **INFECTION** (SARI)a resp , at least mild **INFECTION** , in childre g sublethal IAV **INFECTION** results in i nt of influenza **INFECTION** . Because of othesis , viral **INFECTION** frequently h regulated upon **INFECTION** with IAV , a y to mycoplasma **INFECTION** in cattle ha haviors against **INFECTION** , there shou l and bacterial **INFECTION** through rapi tests confirmed **INFECTION** with SARS-CO lting from MERS **INFECTION** , as well as	0.0322

Test on 100 query words

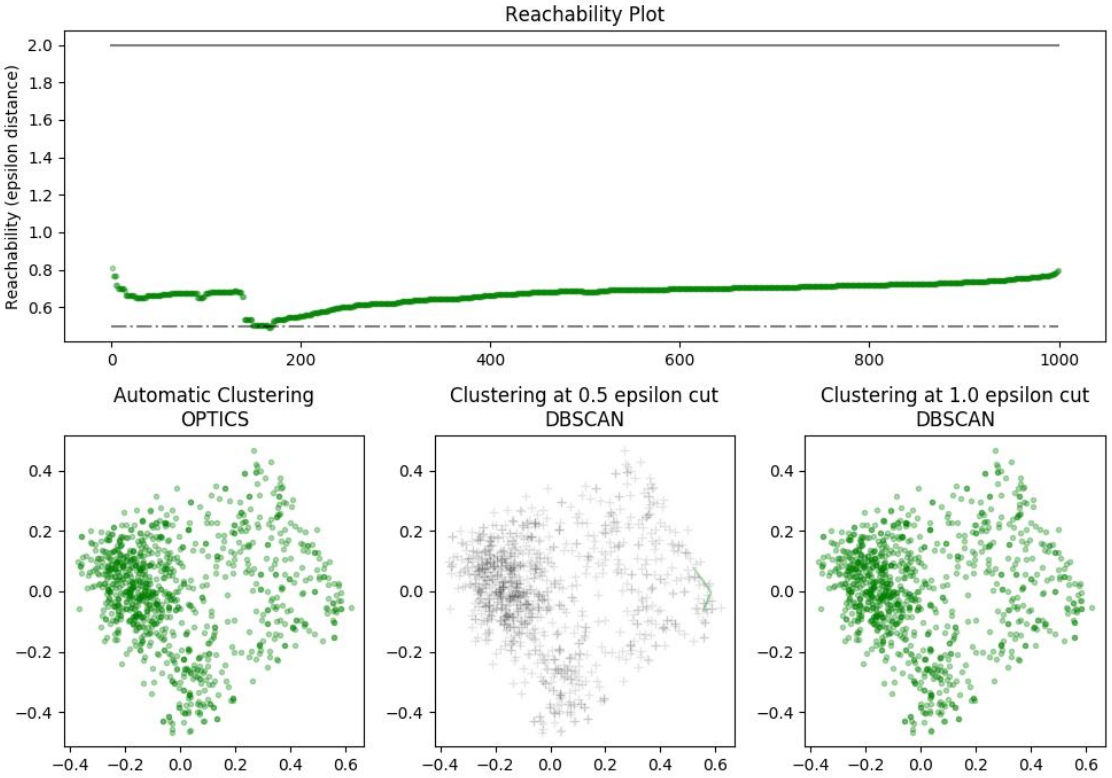


Covid



Assnat

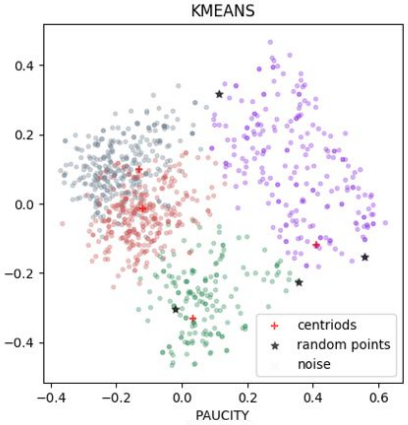
Comparison between clustering techniques

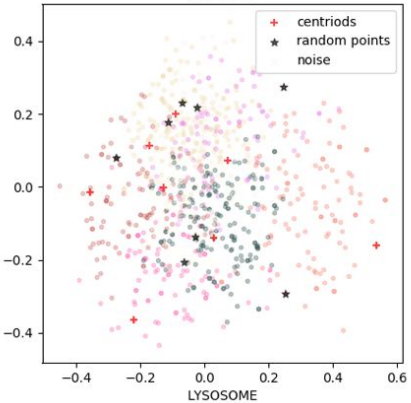


- **Reachability**
  - distance from a data point to a core point
- **OPTICS**
  - all points in one cluster
  - within 1.0 epsilon distance, all points are reachable
- **DBSCAN at 0.5 eps**
  - gray points are regarded as noise
  - no cluster
- **DBSCAN at 1.0 eps**
  - all points in one cluster
  - same result as OPTICS

Note: OPTICS's Xi method can account for different choices of eps in DBSCAN

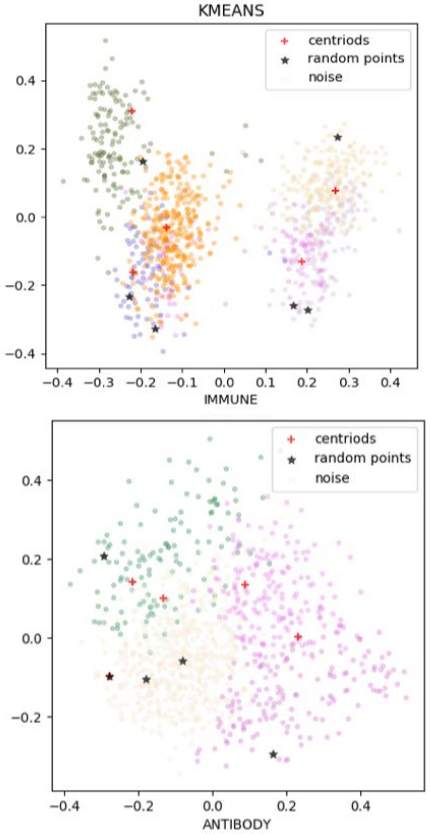
Comparison between clustering techniques





Results	Rouge score
basis , there is a **PAUCITY** of data on the long-term visits Initially the by , **PAUCITY** of data to level values a s , paralleling the **PAUCITY** of melanomas reported in other report in 2 , the **PAUCITY** has of to experimental all	0.0185
blood vessel walls which induce **LYSOSOME** enzyme liberation from neutrophils hydrolyzed by proteases in the **LYSOSOME** , antigen fragments generated by autophagosome subsequently fuses with a **LYSOSOME** , enabling the intra-autophagosomal components directly taken up by the **LYSOSOME** , either through invagination of . to Biol the 17(4) **LYSOSOME** e2007044 . begins with packaging of the substrates into the **LYSOSOME** [107] . The it phagosome directly capture data **LYSOSOME** , a in that instruments the mode of administration of **LYSOSOME** targeted medica tions in order	0.0039

Comparison between clustering techniques

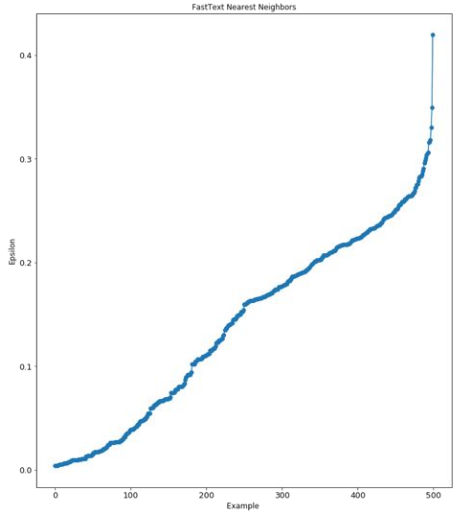


Results	Rouge score
oral administration only elicited an **IMMUNE** response in one of six signaling crosstalk that programs the **IMMUNE** response to virus infection . of autoantigenic material to the **IMMUNE** system but , again , For example , when endogenous **IMMUNE** cells lose their ability to of also IBV induce sequences **IMMUNE** all Asian countries excluding China	0.0667
of cattle induce a strong **ANTIBODY** response , the time delay in our view . Once **ANTIBODY** coating proceeds beyond a critical risk passive of transfer = **ANTIBODY** vaccinated a at protective vaccinee unvaccinated with 1:3000 dilution of anti-His-tag **ANTIBODY** (Invitrogen) in PBS-T antibody cDNA . The obtained **ANTIBODY** cDNA is inserted into expression	0.0

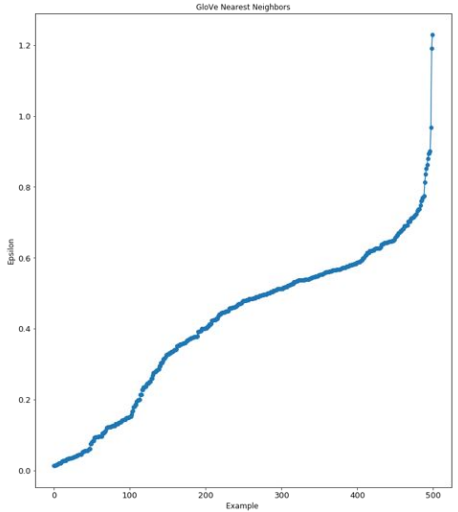
Note: KMeans + Sbert + PCA dim=2 + Window size = 5

Nearest Neighbors' Distance Distributions

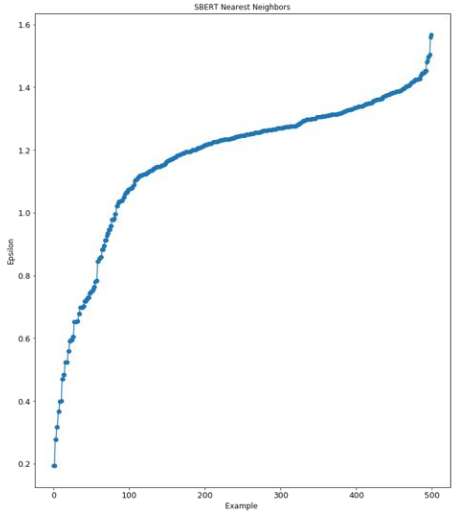
fastText



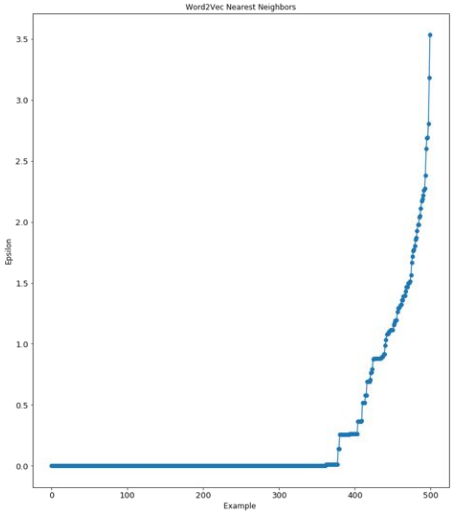
Glove



Sbert

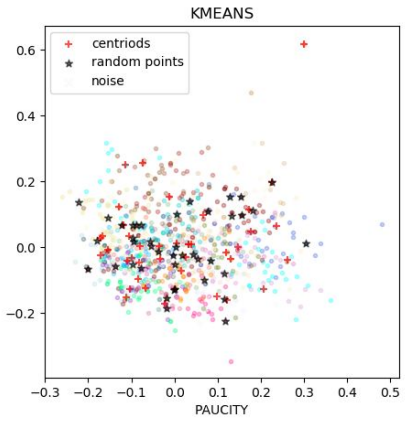


word2vec

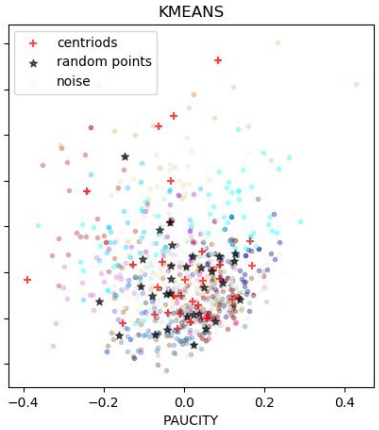


Clustering on Different Features

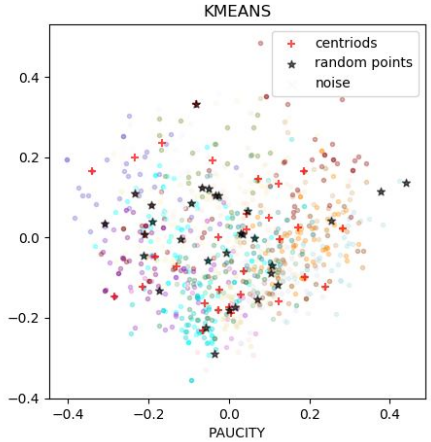
fastText



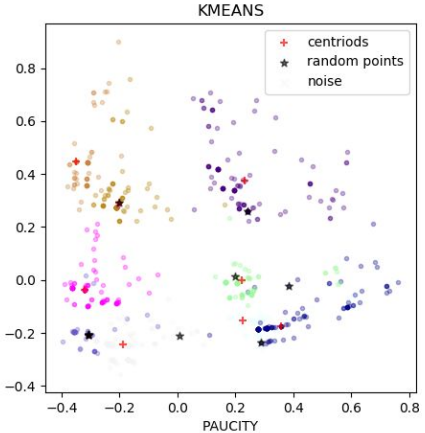
Glove



Sbert



word2vec



Fundamental Differences in Features

Differences in Embedding Methodology:

- How the context feature embeddings (fastText, Glove, sbert, and word2vec) are derived for each method are different.

Differences in Training/Pre-training and Ramifications:

- Word2vec was trained on the corpus while other models were not.

Issues with Using Rouge to Compare Methods:

- Rouge score alone may not really capture the difference in these methods as it is a token similarity based metric.

Challenges and Future Work

- Since ROUGE is an n-gram overlap based approach, it doesn’t take into account the semantic or syntactic level of information contained. There are some other ROUGE variants worth exploring like ROUGE - synonyms and ROUGE - AR.

- Weighting words based on index in the sequence could change the effect of window size on clustering

- Creating our own models for both token-based and context-based vectorization from scratch to be better suited for our corpora domains and task

- Defining desired linguistic differentiation between concordances