Automatic Word Association

CS57-1

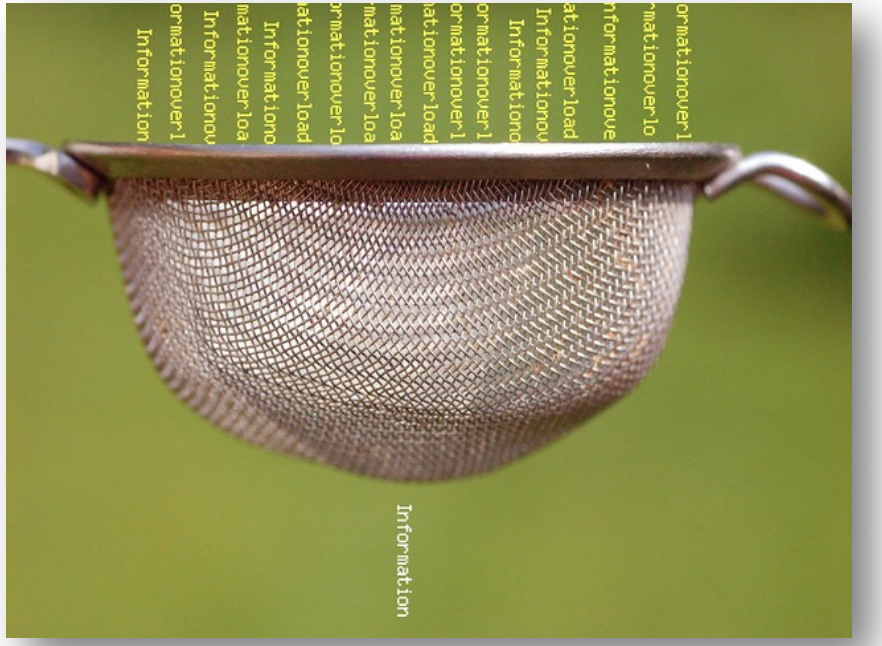
Joseph Tagudin	470542996
Joshua Wildig	430156278
Justin Ho	480450388
Soyeon Kim	470484456
Travis Stenborg	490385087

Background and Aim

EzyReviewer: medical information retrieval system.

Active area of research:

- Wang, H., Zhang, Q., & Yuan, J. (2017). Semantically Enhanced Medical Information Retrieval System: A Tensor Factorization Based Approach



AIM

Implement synonym finding models specifically word embedding

Evaluate each and choose the best.

Compare these models against the client's.

Word Embedding

- Family of techniques
- Maps words to dense vectors
- Builds a simple language model
- Similar words map to similar vectors

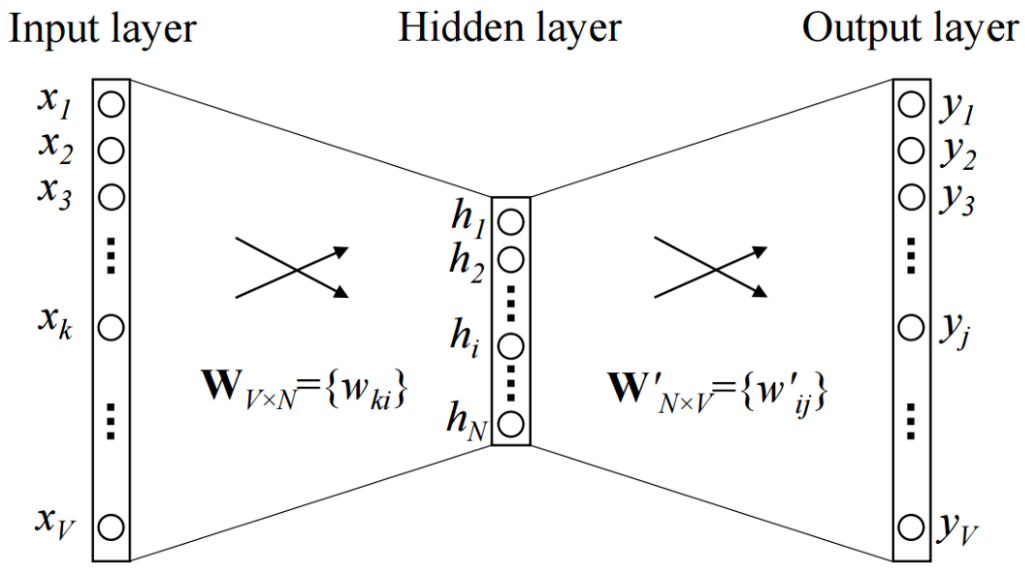
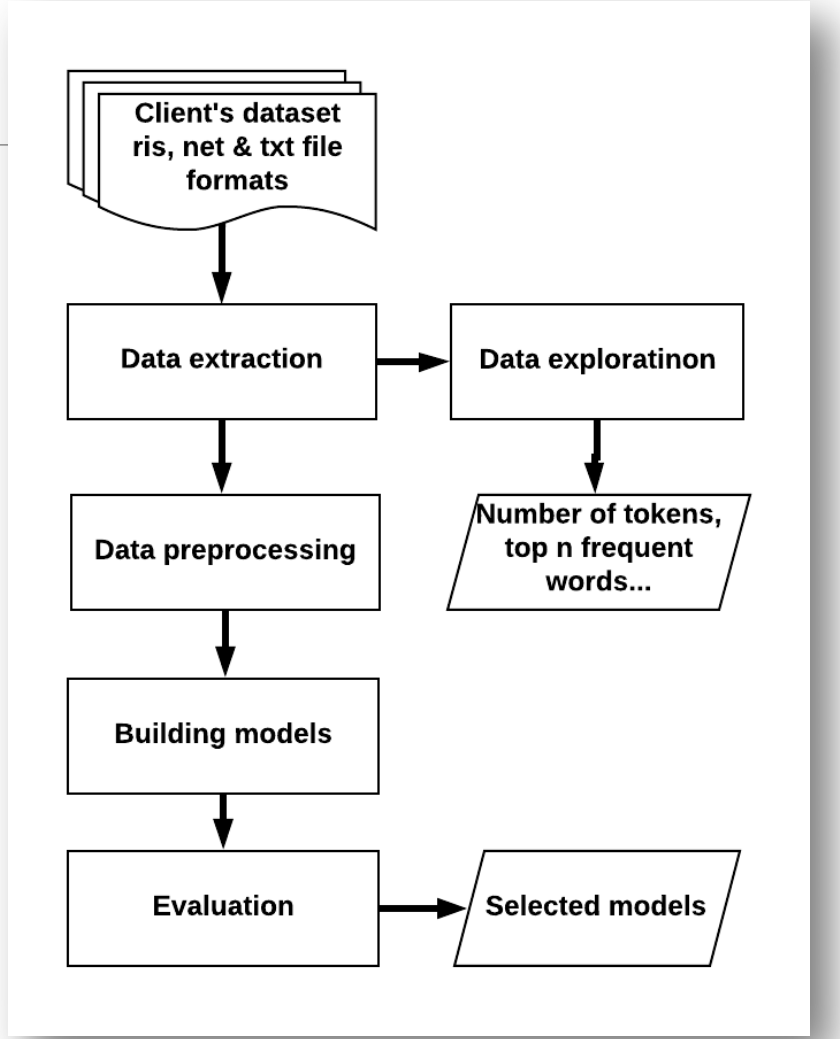


Figure 1: A simple CBOW model with only one word in the context

https://towardsdatascience.com/introduction-to-word-embedding-and-word2vec-652d0c2060fa

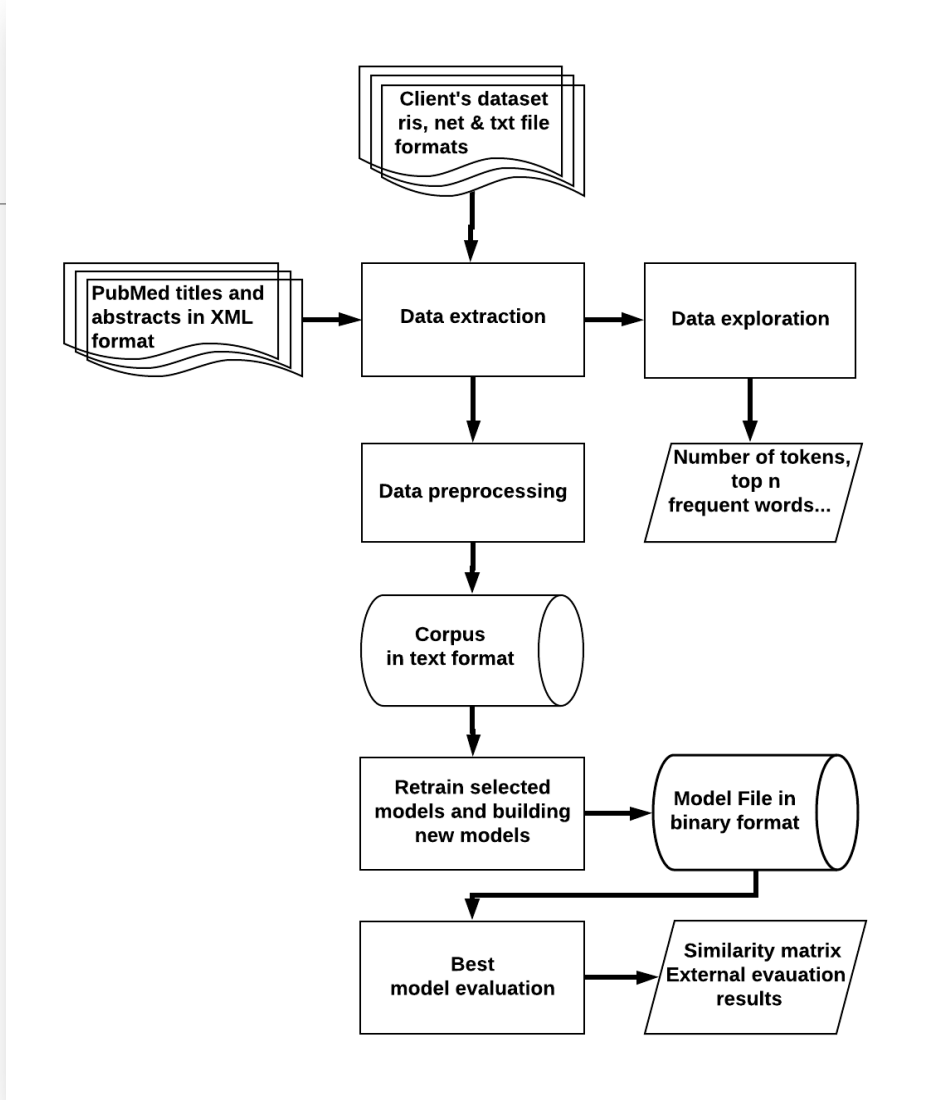
Methodology (1)

- Tools: BeautifulSoup, Gensim, GloVe, matplotlib, NLTK, Pandas, RISparser, SciPy, Seaborn
- Digits replacement
- Removal of stopwords, miscellaneous characters, punctuation and HTML tags
- Tokenisation
- Word2vec and fastText using skip-gram architecture, and GloVe. Used different sizes, dimensions, and training iterations/epochs
- Evaluation
  - Human judgement dataset
  - Client supplied
- Selected models were retrained in stage 2



Methodology (2)

- Tools: as per stage 1 + sklearn + WordCloud packages
- Removal of URLs, white space and unusual leading/trailing characters.
- Lemmatisation
- Embedding models:
  - Word2vec in Skip-gram and CBOW
  - fastText in Skip-gram and CBOW
- Evaluation
  - Human prepared datasets
  - Client supplied
  - External experts



Training Datasets

Evaluation Datasets

Datasets	Term Pairs	Pairs Used (Stage 1)	Pairs Used (Stage 2)
Pedersen*	30	24	29
Hliaoutaki*	34	27	28
MayoSRS*	101	57	89
UMNSRS*	566	151	400
Bio-SimVerb*	1,000	468	960
Bio-SimLex*	988	736	981
Client's Keywords			14,370

*human prepared datasets

Sample from Hilaoutaki

Term 1	Term 2	Physician
Migraine	Headache	0.718
Myocardial Ischemia	Myocardial Infarction	0.750
Sarcoidosis	Tuberculosis	0.406
Amino Acid Sequence	Anti Bacterial Agents	0.156

Results Stage 1: Pearson's r score

Benchmark Models	Human Judgement Datasets			
	D1	D2	D3	D4
EHR20	0.46	0.43	0.12	0.27
EHR60	0.58	0.58	0.45	0.26
EHR100	0.64	0.64	0.47	0.43
PM20	0.41	0.13	0.15	0.13
PM60	0.65	0.65	0.27	0.39
PM100	0.42	0.42	0.19	0.38
Wiki50	0.35	0.35	-0.06	0.17
Wiki100	0.44	0.44	0.01	0.14
GoogleNews	0.48	0.48	0.05	0.29

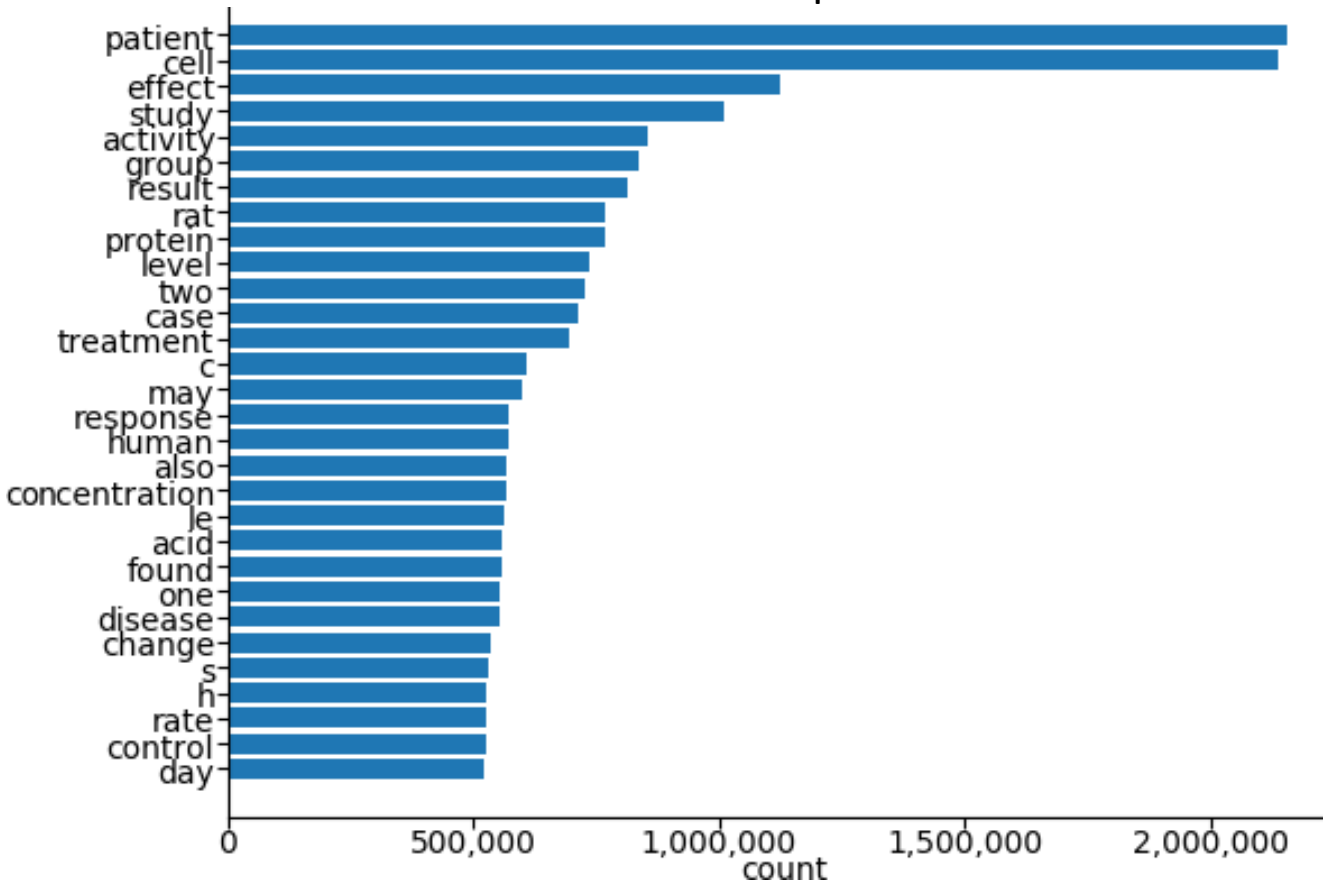
D1 – Pedersen, **D2** – Hliaoutaki, **D3** – MayoSRS
D4 – UMNSRS, **D5** – Bio-SimVerb, **D6** – Bio-SimLex

Our Trained Models	Human Judgement Datasets					
	D1	D2	D3	D4	D5	D6
GloVe200.30	0.48	0.04	0.11	0.14	0.23	0.43
GloVe200.40	0.50	0.08	0.08	0.19	0.25	0.44
Word2vec200.30	0.69	0.29	0.26	0.44	0.37	0.55
Word2vec200.40	0.63	0.30	0.28	0.46	0.37	0.54
fastText200.30	0.78	0.29	0.28	0.54	0.42	0.55
fastText200.40	0.74	0.31	0.31	0.53	0.40	0.54

- Pearson's Correlation: trained models vs Physician (Medical Expert) Word pairs
- Best performing: fastText dimension 200, trained over 30 iterations
- Score increases as number word pairs is increased, this is demonstrated by all the trained models correlated on both Bio-SimVerb and Bio-SimLex datasets

Data Analysis Stage 2:

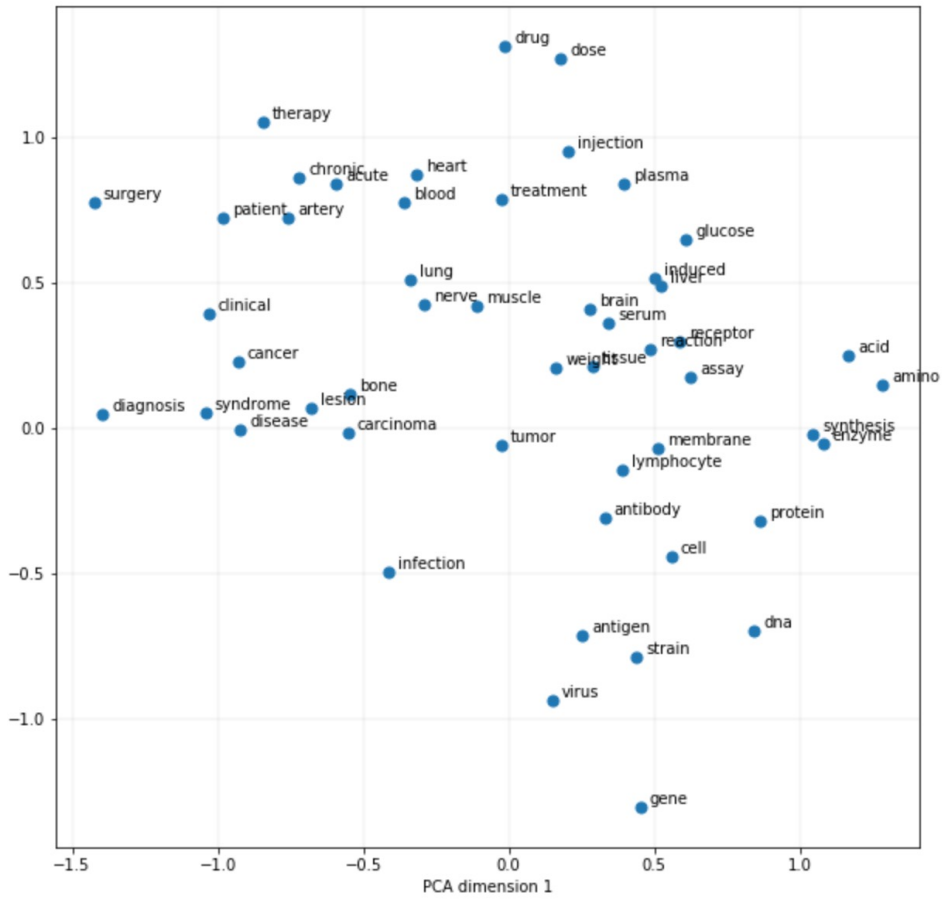
Bar chart of most frequent words



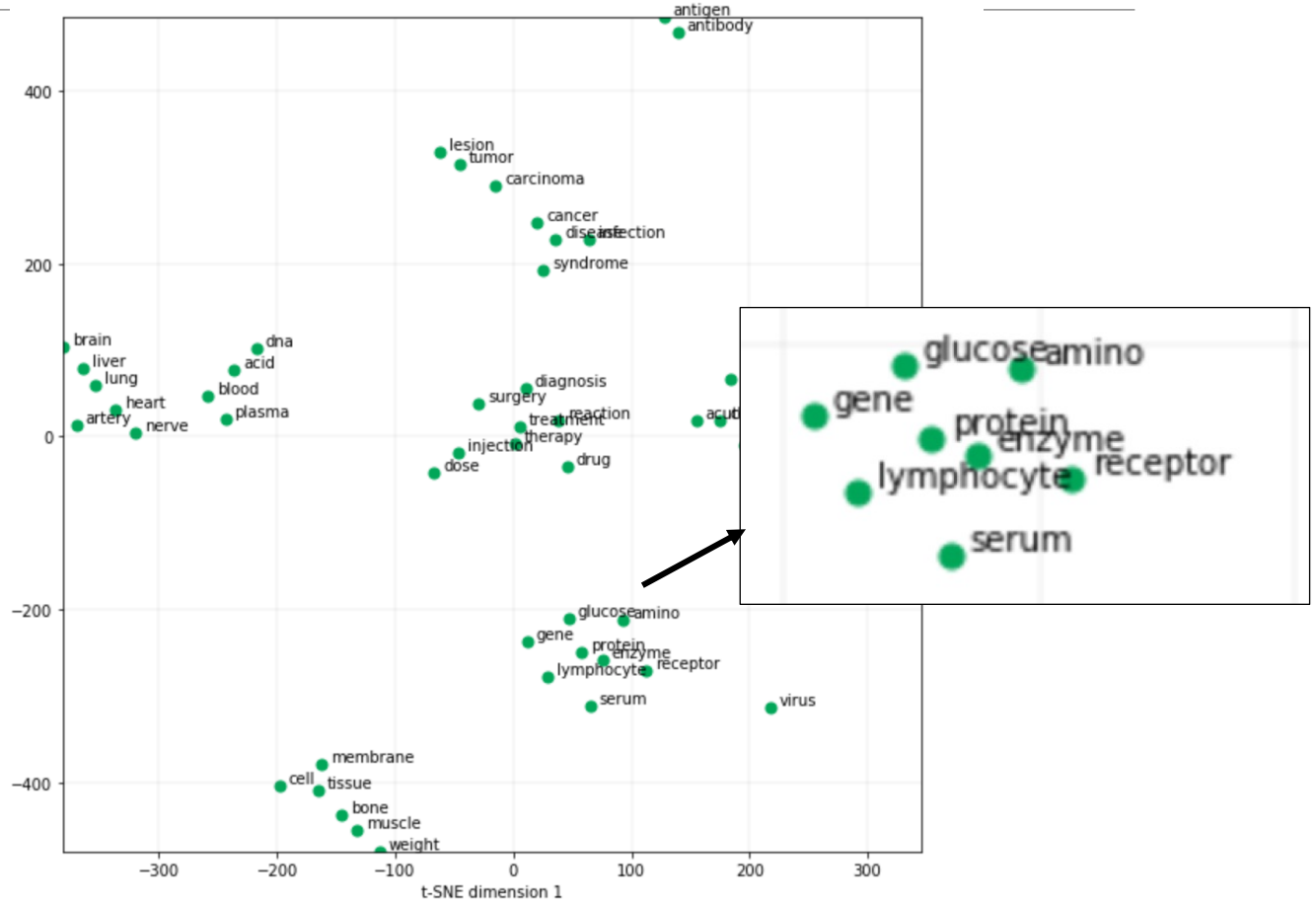
WordCloud of most frequent words

Data Analysis Stage 2: Qualitative Plots

Word2vec Skip-gram 200 dimensions trained over 10 iterations

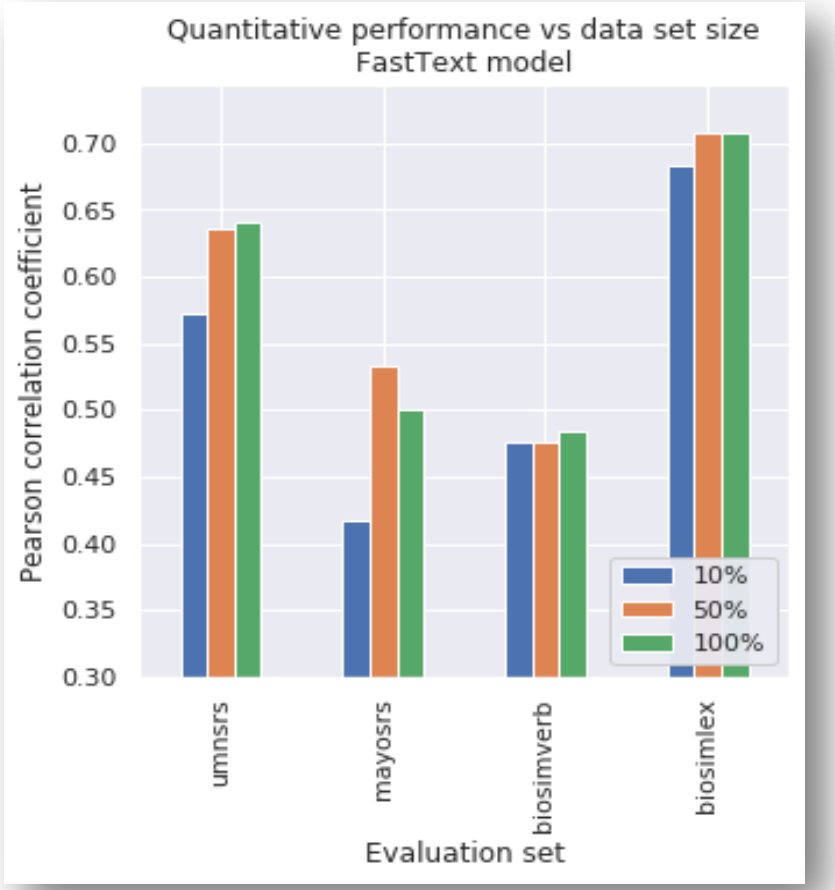


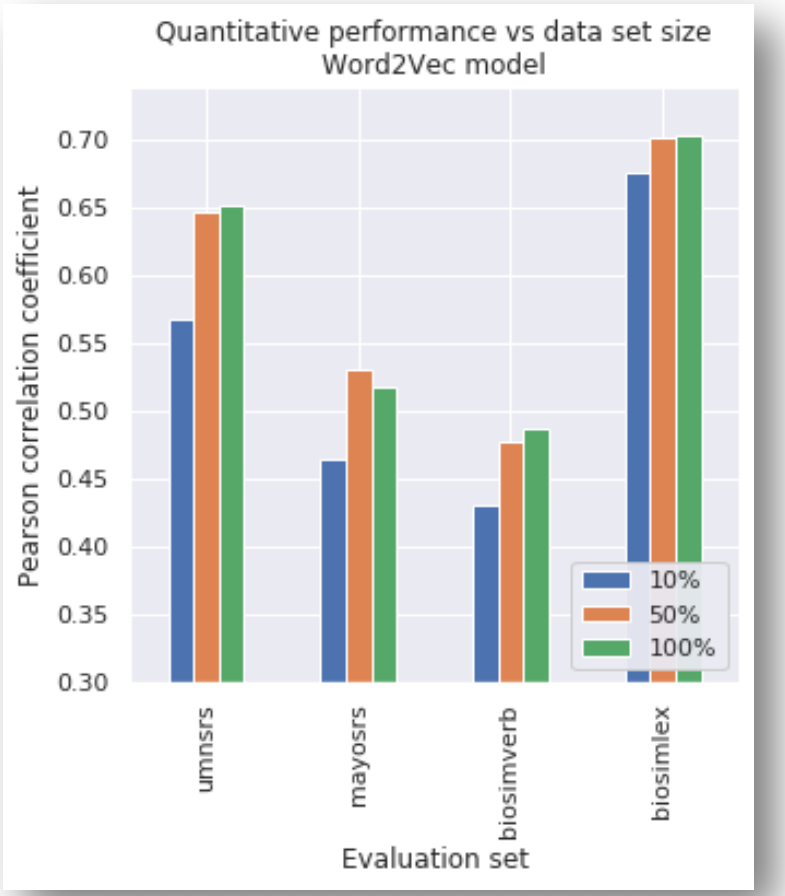
Principal Component Analysis



t-Distributed Stochastic Neighbor Embedding

Results Stage 2: training size impact

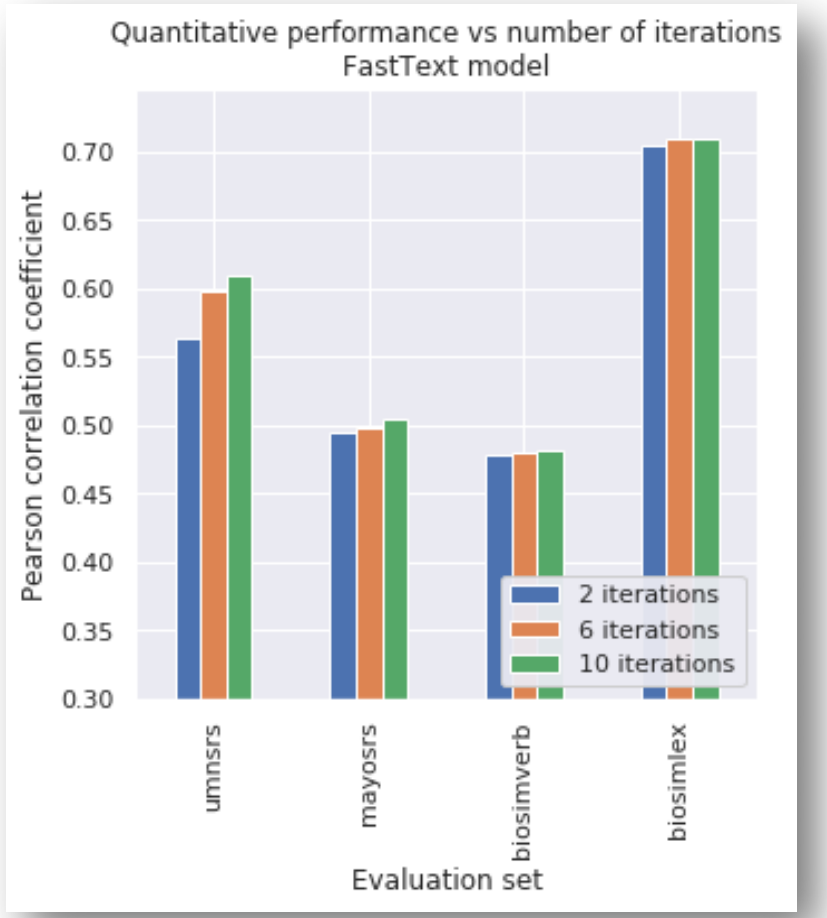


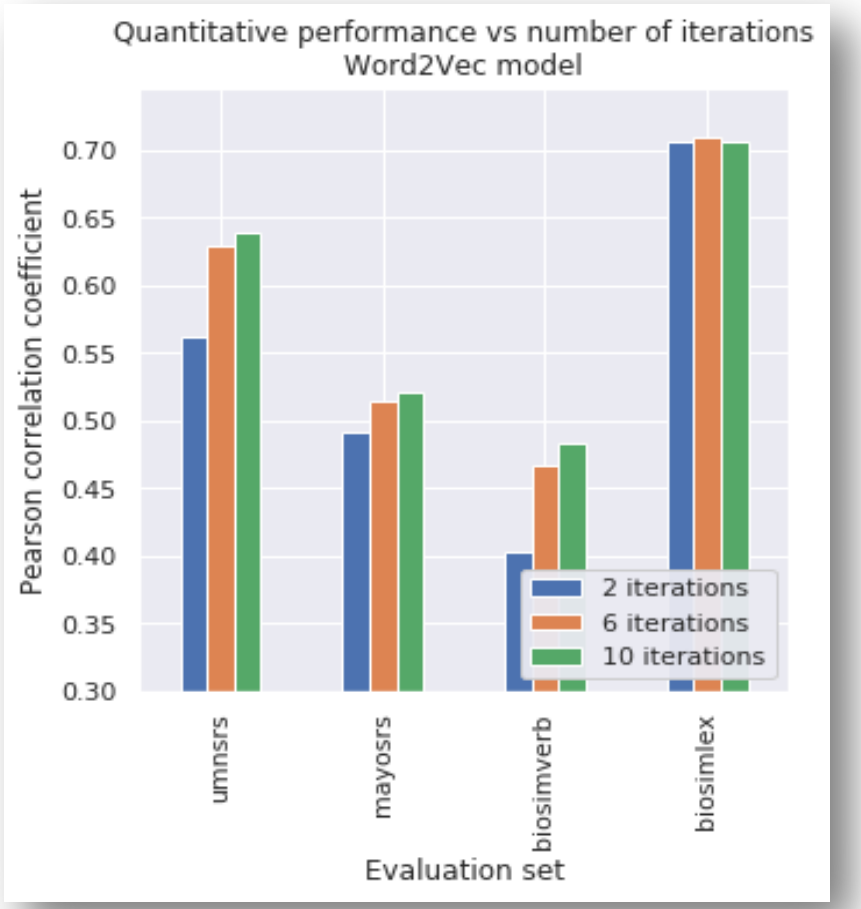


- Word2vec outperforms fastText on 100% of corpus
- fastText outperforms on smaller training data
- Rate of improvement in accuracy is not as great between 50% to 100% => diminishing returns

Results Stage 2: number of iterations impact

- FastText and Word2Vec implemented using Skip-Gram architecture
- Increasing number of iterations tends to have greater improvement in accuracy for Word2vec than fastText





Results Stage 2:
Comparison of synonym sets (Client's vs Ours)

Calculation of similarity metric between results of client and our results

$$W_i = \text{'mofetil'}$$
$$k = 10$$
$$C_i = \text{similar}_{\text{client}}(W_i) = \{\text{'mycophenolate'}, \text{'tacrolimus'}, \text{'cyclosporine'}\}$$
$$M_i = \text{similar}_{\text{model}}(W_i)_{1..k} = \{\text{'mycophenolate'}, \text{'tacrolimus'}, \text{'vepa-m'}, ...\}$$
$$s_i = \frac{|C_i \cap M_i|}{|C_i|} = \frac{2}{3}$$
$$\text{score} = \frac{1}{N} \sum_i s_i$$

- Synonym sets significantly different to client's
- Similarity deteriorates with additional data

Stage 1

Model	Iterations	Similarity Score
fastText Skip-gram	30	0.16

Stage 2

Model	Iterations	Similarity Score
fastText Skip-gram	10	0.044
Word2vec Skip-gram	10	0.098
fastText CBOW	10	0.031
Word2vec CBOW	10	0.101

Results (Stage 2): Client's comparison

Medical Synonyms Survey

This form lists 120 keywords, each with two proposed synonym sets.
For each keyword, select the best synonym set.

* Synonyms are words with the same or very similar meaning.
* Some sets contain rare, contradictory or even nonsense words. Nonetheless, please indicate your opinion of the best overall set in each case.

* Required

thymidine *

- ○ tdr, h-thymidine, dthd
- ○ mpa, ata, coenzymes

vein *

- ○ venous, artery, jugular, saphenous, cava
- ○ deep, right, venous, artery, inserted

- 4 evaluators with medical backgrounds
- 119 keywords and associated synonym sets
- Individual results: 93, 91, 88, 98
- 75% (89 out of 119) preferred synonym sets were from our model.
- Assuming binomial distribution, p-value = 9.29e-09.
- Highly statistically significant result.

Discussion

- Discussion/Evaluation of own work
  - Data issues: duplicate data, data cleaning
  - Data bias: more American terms
  - Lemmatiser: systematic bias eg. abbreviations of American states. RI: Rhode Island, but could also mean respiratory index
- Rationalise why a particular method/model would outperform
  - CBOW vs Skip-gram
  - More data
  - Similarity measures: Jaccard vs Cosine

Limitations

- Numbers are hashed: some loss of meaning eg. Genomics and Proteomics, H1N1 (swine flu) and H5N1
- Additional hyper-parameter tuning
- Additional pre-processing
- Google Collaboratory environment limitations: time and memory
- Evaluation bias: human bias synonymous terms versus associated terms, human behaviour in completing survey (eg. filling in first/last option when in doubt)
- Out of Vocabulary words

Conclusion

- Importance of preliminary data analysis
- Breadth (vocab size) and depth of vocabulary is important in training these models
- Bias can be created and removed
  - pre-processing (lemmatisation)

  evaluation

- Future work and improvements
  - Improved preprocessing: Hyphenated synomyns, Non-alphabetic tokens, Non-alphabetic prefixes and suffixes, removal of DOI (digital reference identifier)
  - More training data: more Pubmed, Embase (more Euro-centric), include articles not just abstracts/titles
  - Incorporate other languages
  - Higher computing power: eg cloud such as AWS, Google, Artemis, Azure
  - Compound words and phrases
  - Greater volume of medical domain expert evaluators

- Successfully created word similarity matrix as a basis for Automatic Medical Word Association
  - Reasonable level of accuracy when evaluated against "human validated" datasets
  - Generally our synonymous terms are shown to be dissimilar to client
  - Insights can be used to improve EzyReviewer

Questions