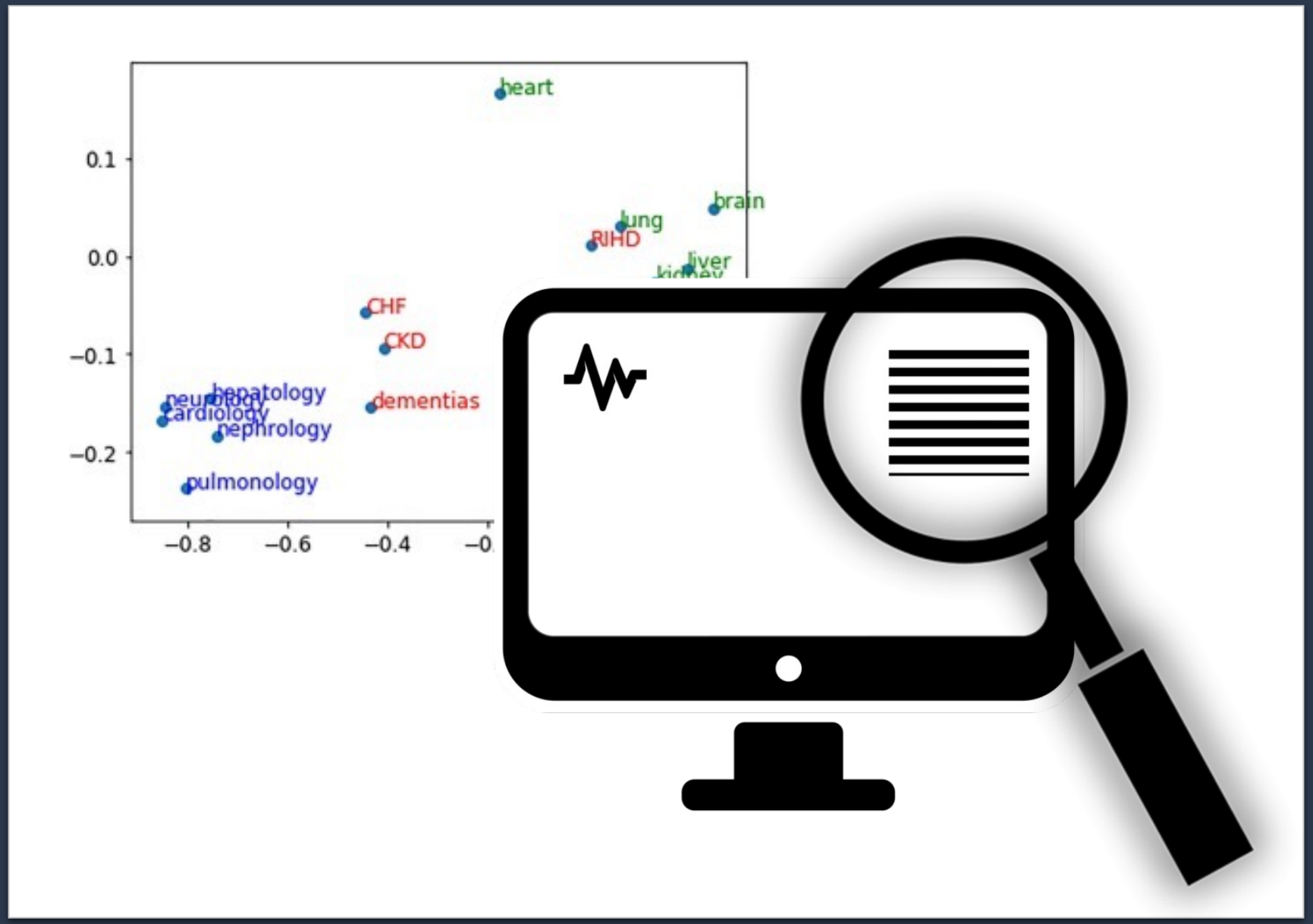
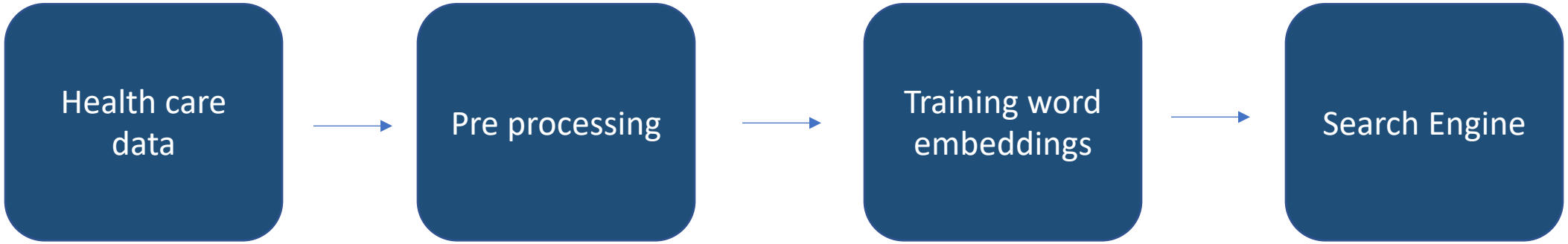


Medical word embeddings *and* search engine

AGENDA

1. Business context and objective
2. Translating into DS approach
3. Importing Data libraries
4. Data Pre processing – Cleaning the text data
5. Exploratory data Analysis
7. Understanding word embeddings
8. Understanding skipgram
9. Understanding Fasttext
10. Model building : skipgram and Fasttext
11. Model embeddings - Similarity and PCA Plot
12. Getting Vectors for each abstract using skipgram and Fasttext
13. Informational retrieval functions
14. Search results evaluation
15. Building streamlit app.py file
16. Running the app and output validation
17. Making production ready code
18. Conclusion

Translating into DS approach



Dataset used :

https://dimensions.figshare.com/articles/dataset/Dimensions_COVID-19_publications_datasets_and_clinical_trials/11961063

What is word embeddings?

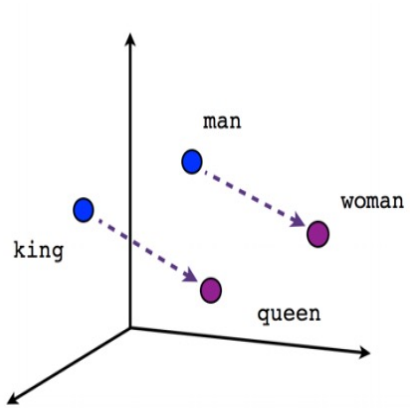
Dimensions

Word vectors				
dog	-0.4	0.37	0.02	-0.34
cat	-0.15	-0.02	-0.23	-0.23
lion	0.19	-0.4	0.35	-0.48
tiger	-0.08	0.31	0.56	0.07
elephant	-0.04	-0.09	0.11	-0.06
cheetah	0.27	-0.28	-0.2	-0.43
monkey	-0.02	-0.67	-0.21	-0.48
rabbit	-0.04	-0.3	-0.18	-0.47
mouse	0.09	-0.46	-0.35	-0.24
rat	0.21	-0.48	-0.56	-0.37

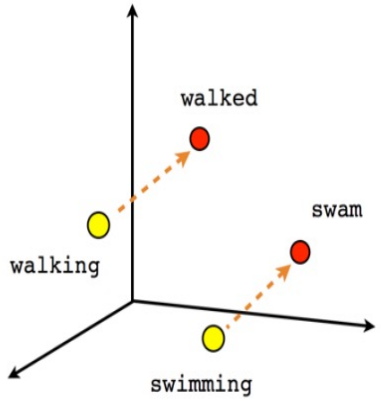
animal
domesticated
pet
fluffy

Word embedding is the feature learning techniques in natural language processing(NLP) where words or phrases from the vocabulary are mapped to vectors of real numbers capturing the **contextual hierarchy**.

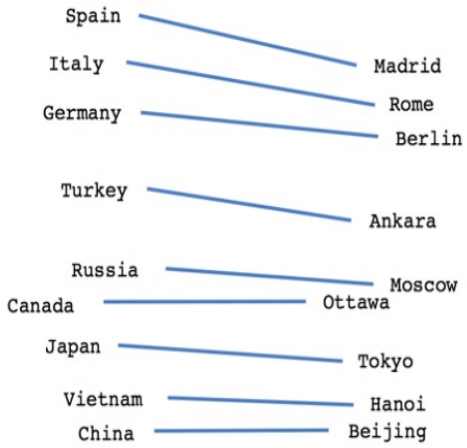
Word embedding is derived from shallow neural networks.



Male-Female



Verb tense



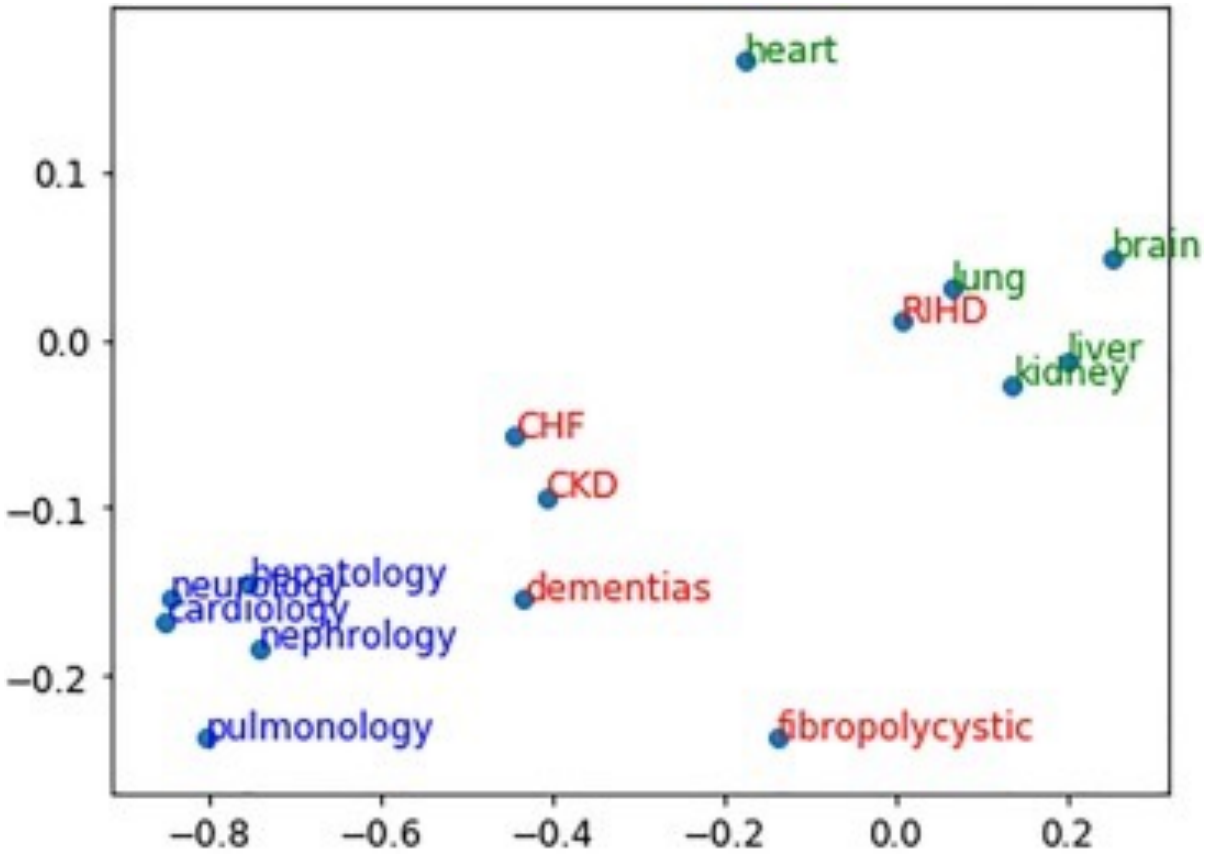
Country-Capital

What are Medical word embeddings?

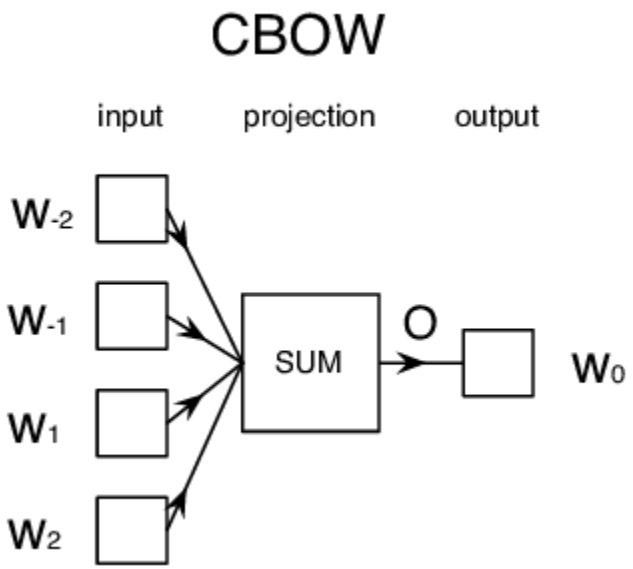
General word embeddings might not perform well enough on all the domains. We need to build domain specific embeddings to get better outcomes.

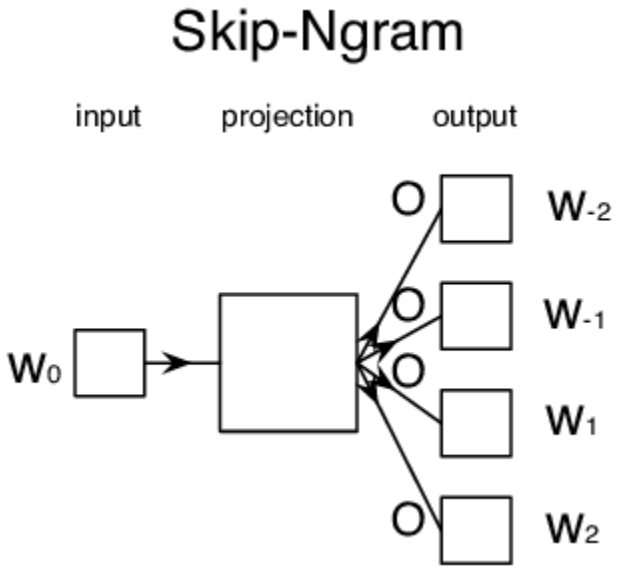
Medical word embeddings are trained purely on medical/healthcare data.

Ex : https://github.com/ncbi-nlp/BioSentVec



Understanding word2vec





In the **CBOW model**, the distributed representations of context (or surrounding words) are combined to predict the word in the middle.

Skipgram tries to predict the source context words (surrounding words) given a target word. It's the reverse of CBOW

https://arxiv.org/abs/1310.4546

Understanding word2vec - skipgram model

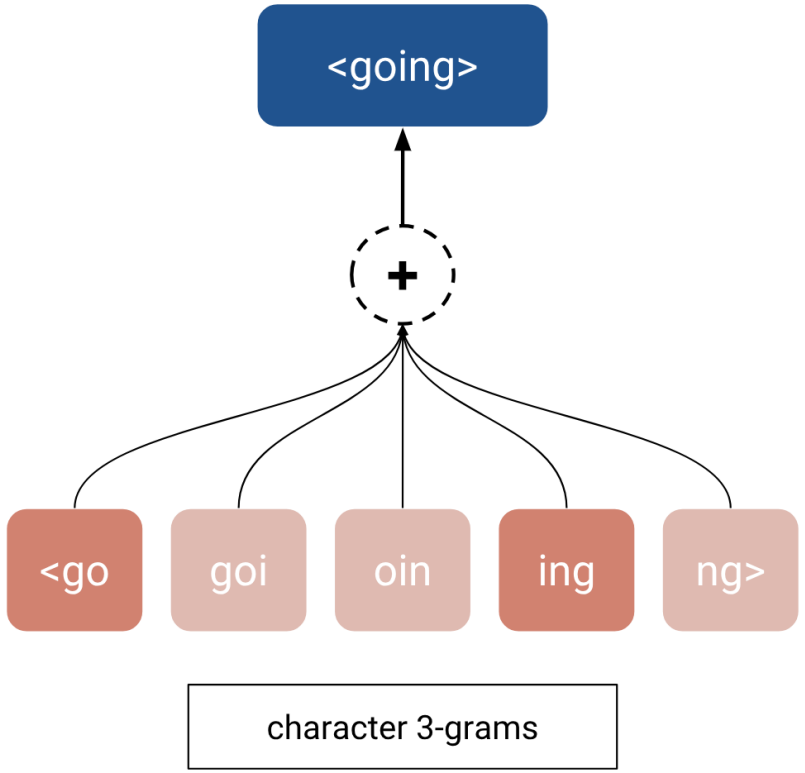
Training Example	Context Word	Target Word
#1	(i, natural)	like
#2	(like, language)	natural
#3	(natural, processing)	language
#4	(language)	processing

Advantages

1.It is unsupervised learning hence can work on any raw text given.
2.It requires less memory comparing with other words to vector representations.
3.works well with a small amount of the training data, represents well even rare words or phrases

Ref : https://thinkinfi.com/word2vec-skip-gram-explained/

Understanding fasttext model



FastText is the improvised version of Word2Vec. Word2Vec basically considers words to build the representation. But fastText takes each character while computing the representation of the word.

Conclusion

- Biomedical pretrained embeddings
- Advanced models like BERT
- Spelling correction techniques