WOMEN WHO CODE DATA SCIENCE

Deep Dive (3)

Text Vectors, Word Embeddings, Word2vec, Glove, Implement with Gensim, Use-cases and overview of BERT

Source: https://www.sas.com/en_us/insights/analytics/what-is-natural-language-processing-nlp.html

WOMEN WHO CODE DATA SCIENCE

Agenda

1. Recap Week 3
2. Text Vectorization & Challenges
3. Word Embeddings
    - Word2vec
    - Stanford GloVe
4. Use-cases
    - Doc Similarity, Question Answering, Information Retrieval
5. State of NLP today - intro to BERT
6. Theory Wrap-up
7. Implement with Google Colab

Recap - Week 3

WOMEN WHO CODE DATA SCIENCE

TF-IDF

Bag-of-Words

Document Similarity - Use Cases

WOMEN WHO CODE DATA SCIENCE

Recap TF-IDF

$$w_{x,y} = tf_{x,y} \times \log\left(\frac{N}{df_x}\right)$$

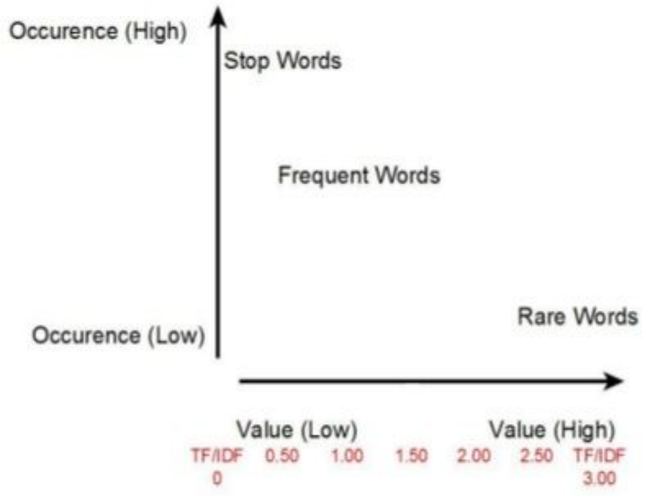
TF-IDF

Term x within document y

$tf_{x,y}$ = frequency of x in y

df_x = number of documents containing x

N = total number of documents

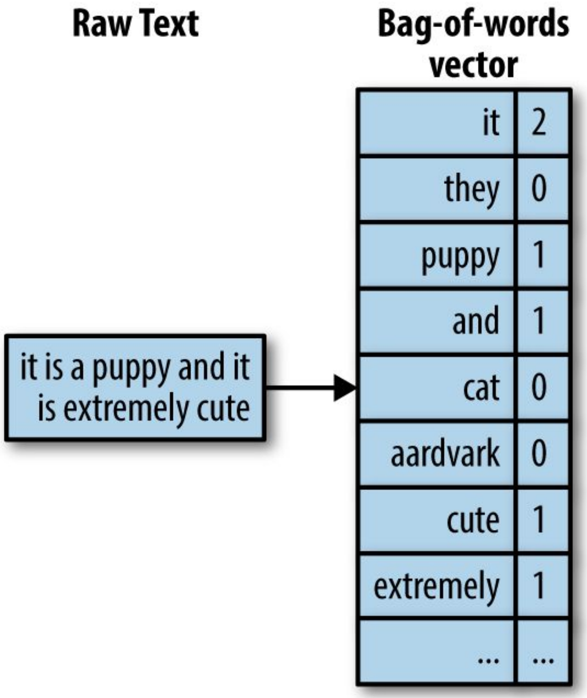


Credit: http://trimc-nlp.blogspot.com/2013/04/tfidf-with-google-n-grams-and-pos-tags.html

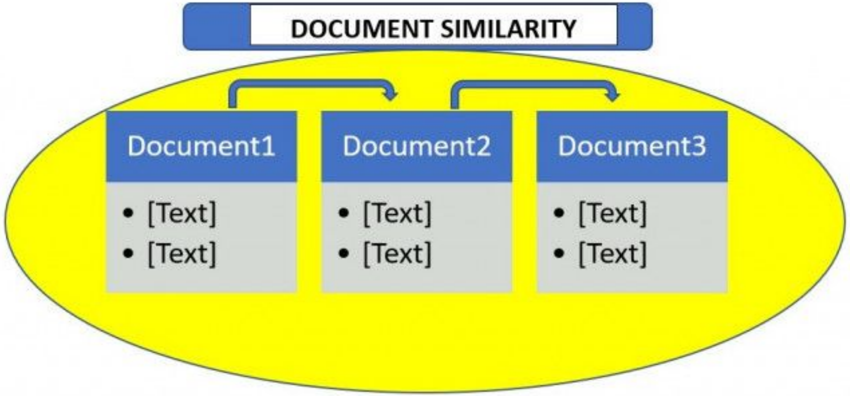
WOMEN WHO CODE DATA SCIENCE

Recap

Bag-of-Words



Recap Document Similarity

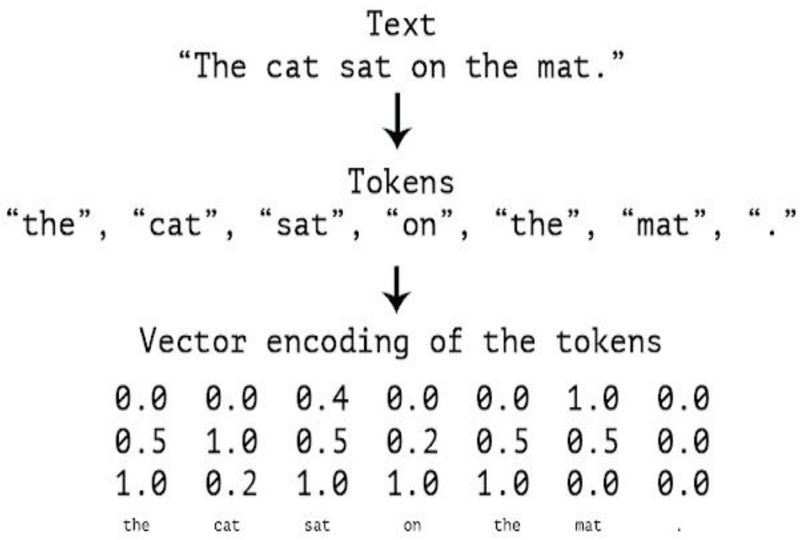


1

What is **Word Embedding**?

Word Vectorization Review

- Vectorization is the process of converting text into numerical representation
- Few techniques available, each one with its own pros and cons
  - simplest encoding techniques do not retain word order
  - fast and intuitive, but the size of document vectors grows quickly with the size of the dictionary
  - optimize dimension but lose in interpretability
- Not only words - sentences, documents etc.

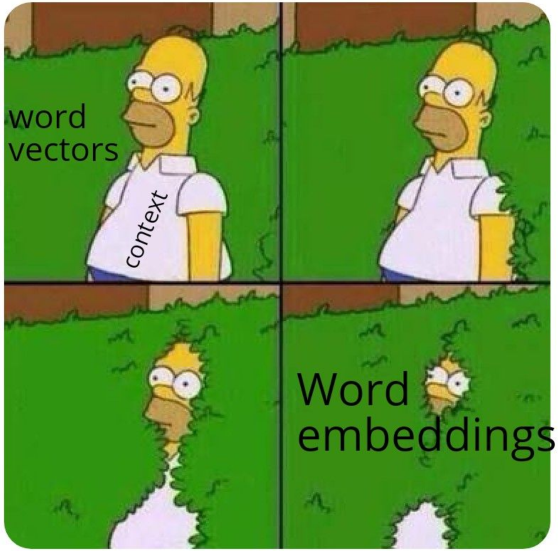


Source: https://www.kdnuggets.com/2019/11/text-encoding-review.html

Word Embeddings

- Embeddings are types of knowledge representation where each textual variable is represented with a vector
    - where words or phrases from the vocabulary are mapped to vectors of real numbers
    - mathematical embedding from a space with many dimensions per word to a continuous vector space with a lower dimension.
- Methods to generate this mapping include:
    - neural networks
    - dimensionality reduction on the word co-occurrence matrix
    - probabilistic models

Word embeddings in a nutshell

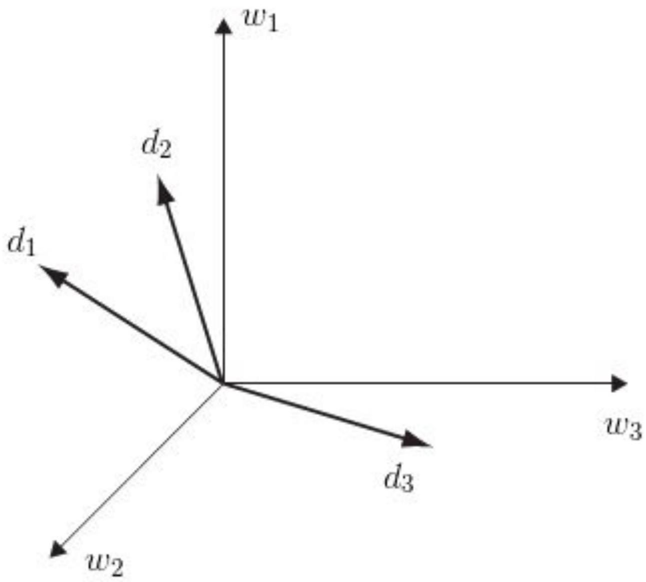


Source: http://mlwiki.org/index.php/Curse_of_Dimensionality

What is a Vector Space

Vector space model is a statistical model for representing text information for -

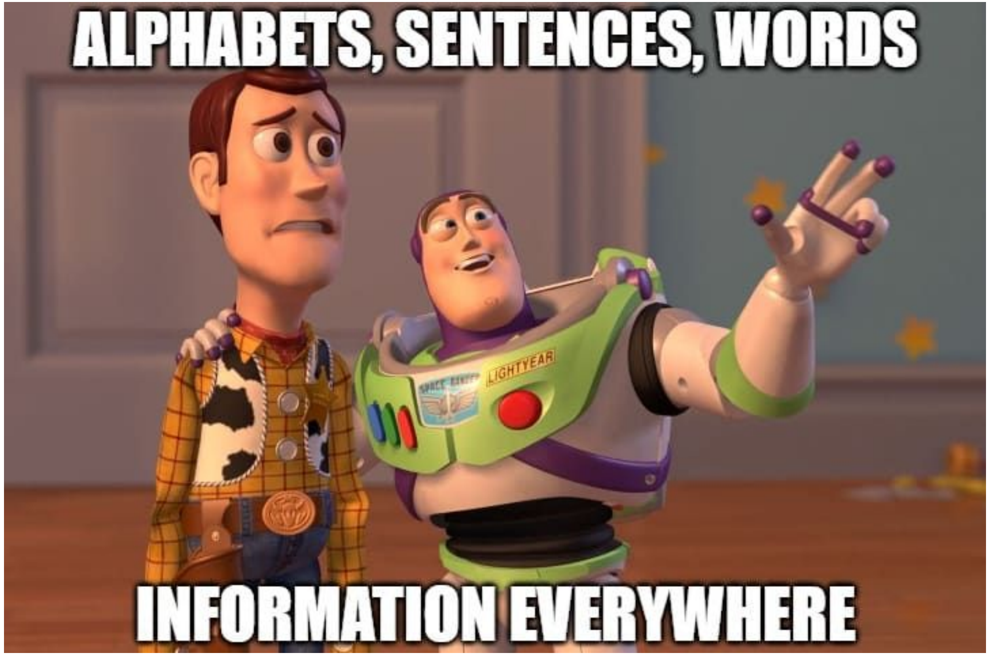
- how many documents contain a term
- what are important terms each document has etc.
- dimensionality - whether vectors are sparse
- e.g. vocabulary size |V|=105, but documents may contain only 500 distinct words
- lexicon of document, word correlations etc.



Source: http://mlwiki.org/index.php/Vector_Space_Models

WOMEN WHO CODE DATA SCIENCE

Google Word2Vec



Word2Vec: Idea

Fundamental idea behind Word2vec?

You shall know a word by the company it keeps — J.R. Firth (1957)

It's all about **Context**

The baseline pre-trained word2vec model has -

- 300 dimensions
- 3 million 'unique' words from google news data in the training corpus

Word2Vec: Context Matters

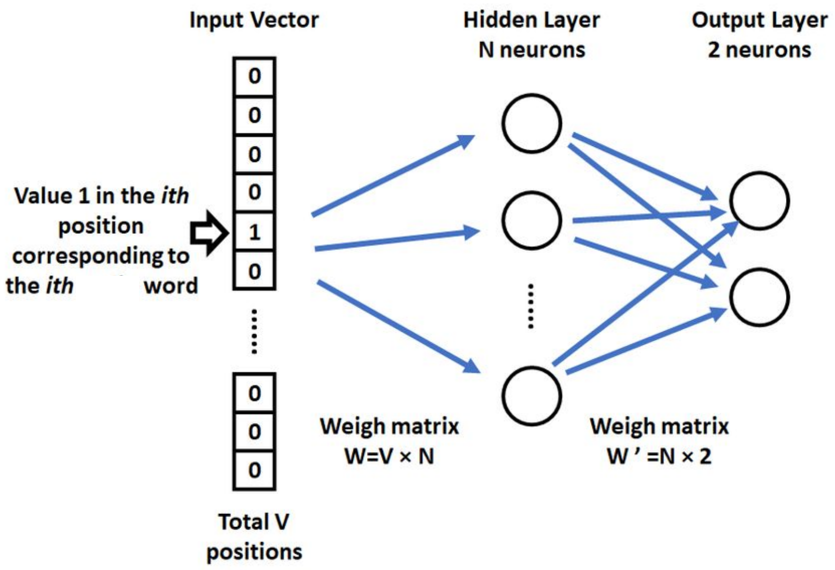
Type of Similarity 1: Semantic Relatedness

Def: some relation between words
Eg: Truck -- Road, Bee -- Honey

Type of Similarity 2: Semantic Similarity

Def: words used in a same way and interchangeable context
Eg: Car -- Auto, Doctor -- Surgeon

Word2Vec: Definition



- A word2vec model is a **feed forward shallow neural network** model with a **single hidden layer**.
- Each word is represented by a vector (Array of numbers based on Embedding Size)
- Word2Vec finds relation (Semantic or Syntactic) between the words which was NOT possible by the Traditional TF-IDF or Frequency based approach
- Transforms the unlabeled raw corpus into labeled data (by mapping the target word to its context word), and learns the representation of words in a classification task

Source: https://www.researchgate.net/figure/NN2Vec-fully-connected-neural-network_fig2_325284755

Source: https://www.kdnuggets.com/2018/08/understanding-language-syntax-and-structure-practitioners-guide-nlp-3.html

Word2Vec: General Process

- When we train a model, each one hot encoded word gets a point in a dimensional space where it learns and groups the words with similar meaning
- Create an Embedding Look up Layer

$$\underbrace{\begin{bmatrix} 0 & 0 & 0 & \boxed{1} & 0 \end{bmatrix}}_{\text{One-hot vector}} \times \underbrace{\begin{bmatrix} 8 & 2 & 1 & 9 \\ 6 & 5 & 4 & 0 \\ 7 & 1 & 6 & 2 \\ \boxed{1 \quad 3 \quad 5 \quad 8} \\ 0 & 4 & 9 & 1 \end{bmatrix}}_{\text{Embedding Weight Matrix}} = \underbrace{\begin{bmatrix} 1 & 3 & 5 & 8 \end{bmatrix}}_{\text{Hidden layer output}}$$

Source: https://towardsdatascience.com/what-the-heck-is-word-embedding-b30f67f01c81

Word2Vec: General Process

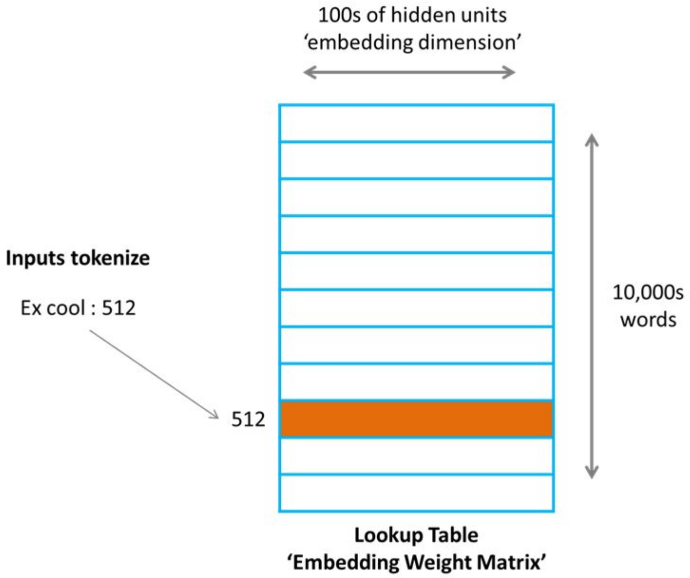
Key Points:

a) The embedding layer is just a hidden layer

b) The lookup table is just a embedding weight matrix

c) The lookup is just a shortcut for matrix multiplication

d) The lookup table is trained just like any weight matrix



Word2vec falls under prediction based embeddings which tends to predict a word in a given context

Source: https://towardsdatascience.com/what-the-heck-is-word-embedding-b30f67f01c81

Word2Vec: Example (1)

Eg 1: Vector of the words similar

1. Cat
2. Dog

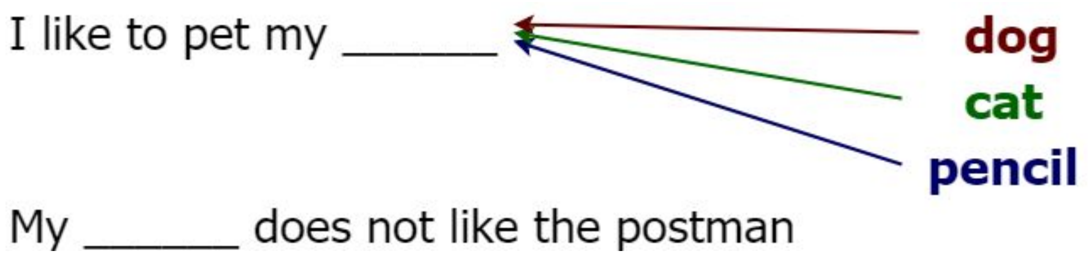
Eg 2: Vector of the words not similar

1. Cat
2. Pencil

<u>Similarity</u> is defined by the frequency of the two words in discussion

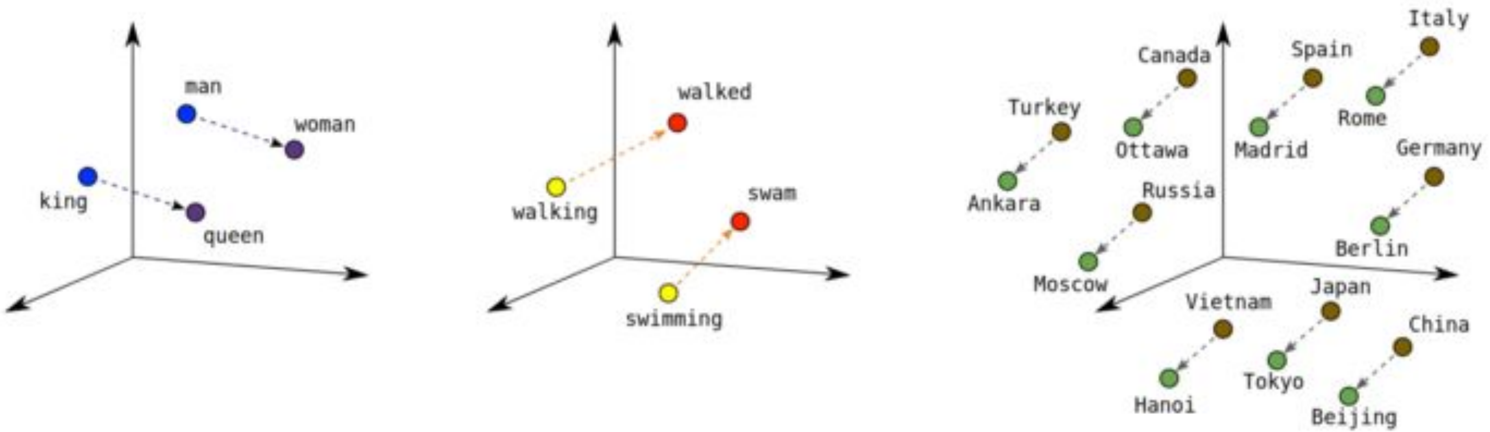
- [cat,dog]
- [cat,pencil]

And how many time they are used in the **<u>same context</u>**



Source: https://medium.com/analytics-vidhya/word2vector-using-gensim-e055d35f1cb4

Word2Vec: Example (2)



Male-Female

Verb Tense

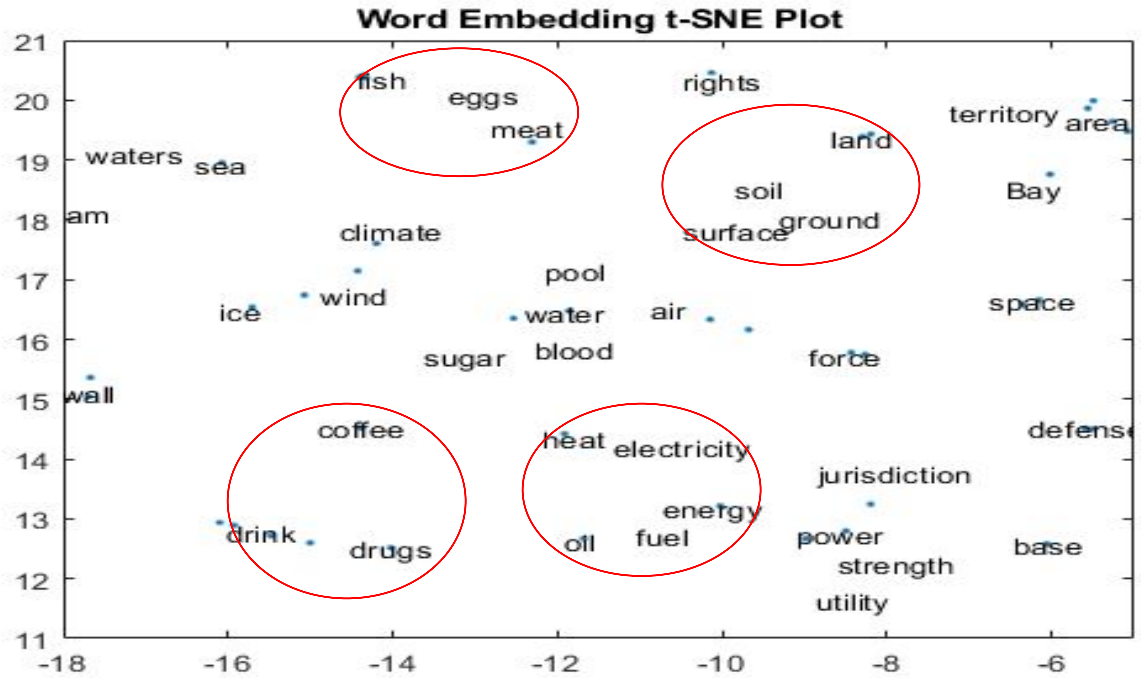
Country-Capital

Word2Vec allows some mathematical operations on vectors

king — man + woman = queen

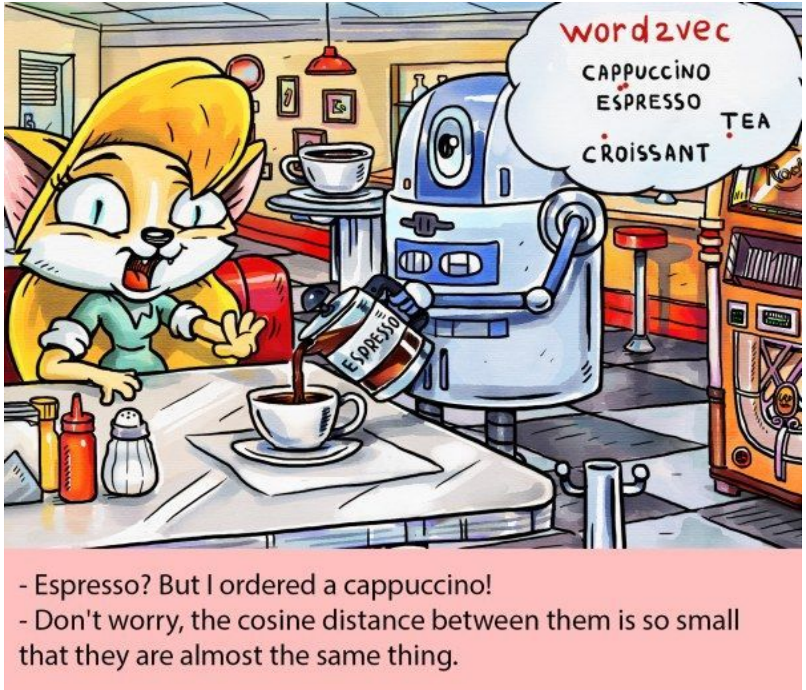
Source: developers.google.com

Word2Vec: Example (3)



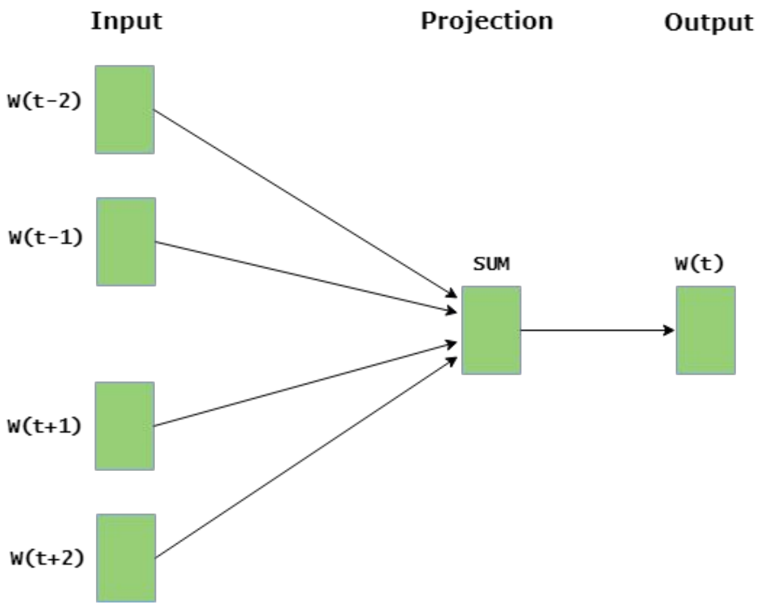
Word2Vec: SOTA Applications

1. Topic Modeling
2. Document Similarity
3. Speech Recognition
4. Chatbots
5. Information Retrieval
6. Machine Translation
7. Question Answering
8. Recommendation Engines



Source: https://twitter.com/hashtag/word2vec

Word2Vec: Architecture (CBOW)

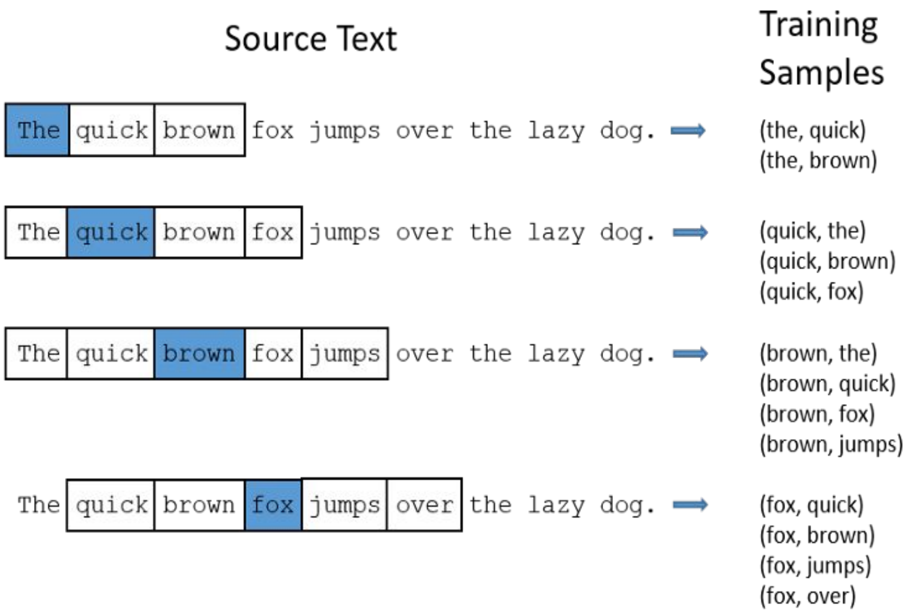


- CBOW - **<u>predict</u>** the current word based on context. Here the *input* will be the **<u>context neighboring</u>** words and *output* will be the ***target word***.
- Based on Window Method, where we have to assign a Window size
- If window size is set to 1. So, 1 word from both the sides of target are being considered. Similarly, in each iteration, window will slides by single stride and our neighbors will keep changing

Source: https://www.geeksforgeeks.org/python-word-embedding-using-word2vec/

Word2Vec: Architecture (CBOW)

WOMEN WHO CODE DATA SCIENCE

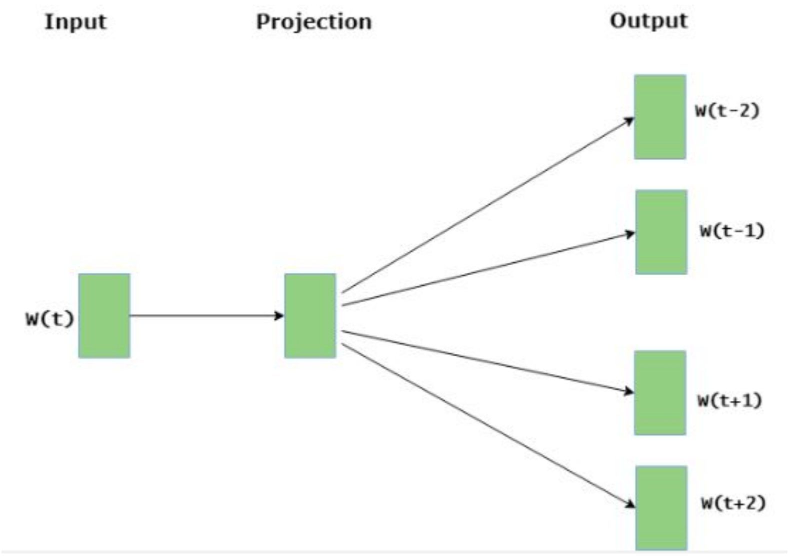


- Model: CBOW
- INPUT Layer: White box content
- TARGET Layer: blue box word
- Window Size: 5

CBOW architecture predicts the current word based on the context

Source: https://hackernoon.com/stepping-into-nlp-word2vec-with-gensim-e7c54d9a450a

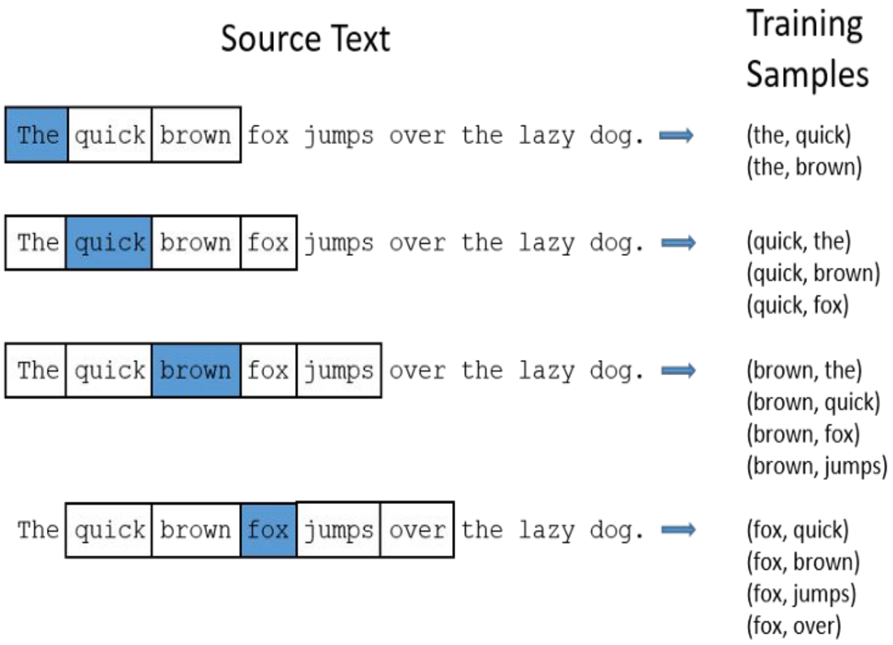
Word2Vec: Architecture (skip-grams)



- Skip gram predicts the **<u>surrounding context words</u>** within specific window **<u>given current word</u>**
- The input layer contains the current word
- The output layer contains the context words
- The hidden layer contains the number of dimensions in which we want to represent current word present at the input layer

Source: https://www.geeksforgeeks.org/python-word-embedding-using-word2vec/

Word2Vec: Architecture (skip-grams)

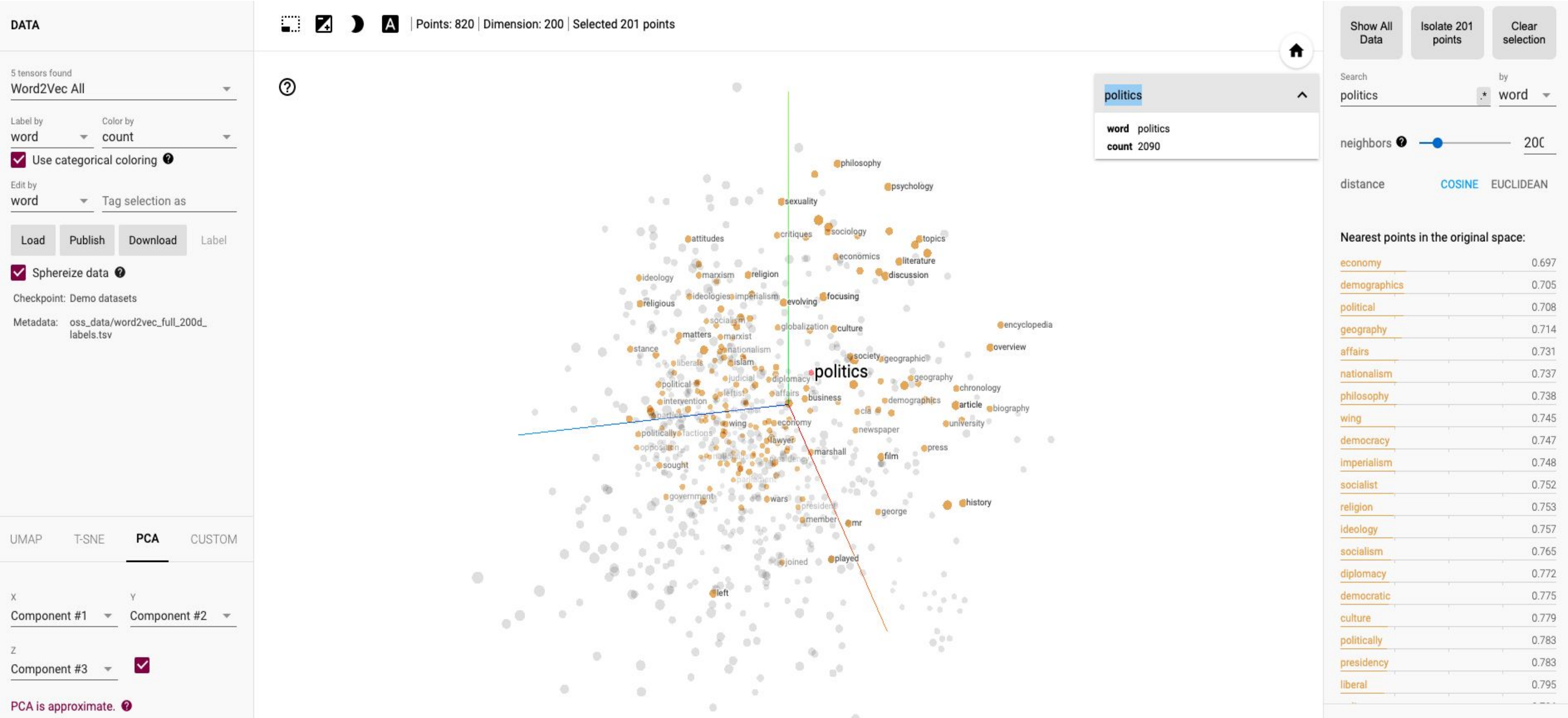


- Model: Skip Gram
- INPUT Layer: Blue box word
- TARGET Layer: White box content
- Window Size: 5

Skip-gram architecture predicts surrounding words given the current word

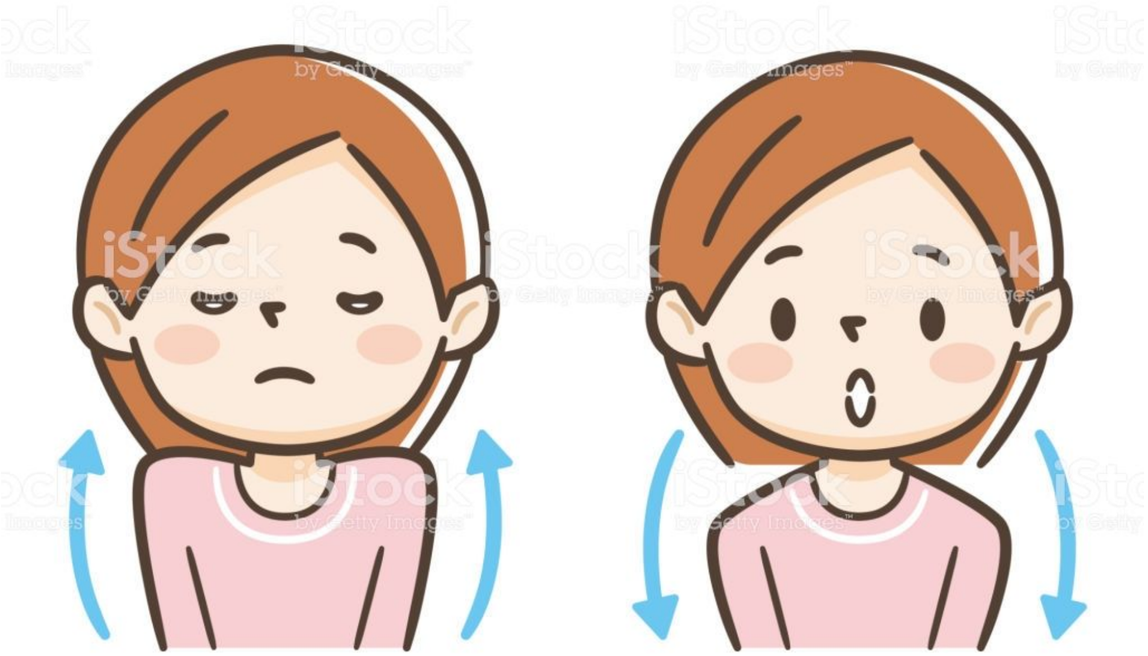
Source: https://hackernoon.com/stepping-into-nlp-word2vec-with-gensim-e7c54d9a450a

Embedding Projector (projector.tensorflow.org)



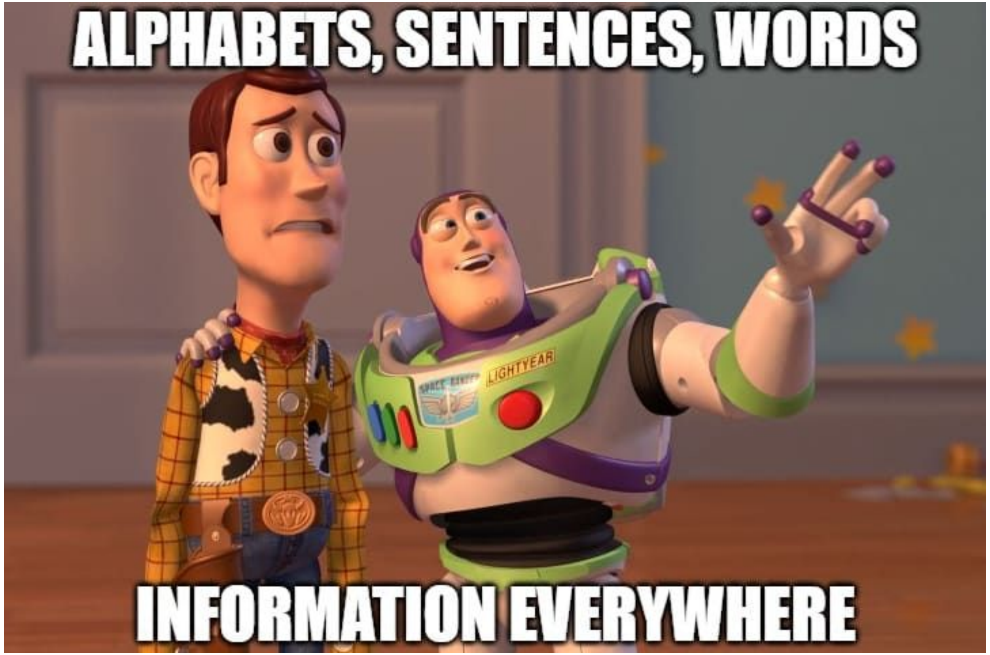
Let’s take a few deep breaths now!

WOMEN WHO CODE DATA SCIENCE



WOMEN WHO CODE DATA SCIENCE

Stanford gloVe



gloVe: Idea

Fundamental idea behind gloVe?

Word2vec: *relies only on **local information** of language. The semantics learnt for a given word, is only affected by the surrounding words*

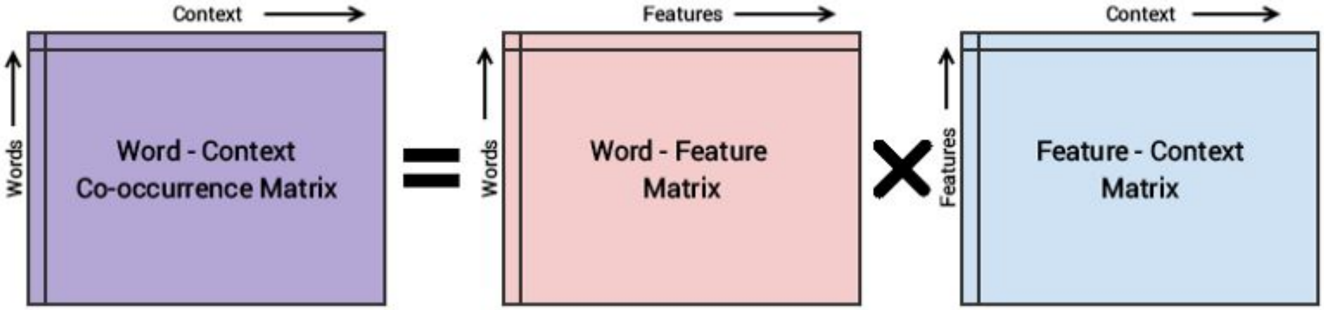
Eg: “The cat sat on the mat”
Con: is “the” a special context of the words “cat” and “mat”?

gloVe: *You can derive semantic relationships between words from the co-occurrence matrix.*

It’s all about **Global** **Context**. GloVe stands for **“Global Vectors”**

gloVe: Definition

- Unsupervised learning algorithm for obtaining vector representations for words.
- Training is performed on aggregated global word-word co-occurrence statistics from a corpus
- Resulting representations showcase linear substructures of the word vector space



gloVe: Process

The relationship of these words can be examined by studying the ratio of their co-occurrence probabilities with various probe words, k.

Eg:
P(k|w): probability that the **word k** appears in the context of ***word w***

Consider a word strongly related to ***ice***, but not to ***steam***, such as **solid**

P(solid | ice) will be relatively high
P(solid | steam) will be relatively low

Ratio of P(solid | ice) / P(solid | steam) will be large

gloVe: Process

Consider the word **gas** that is related to ***steam*** but <u>not</u> to ***ice***,

Ratio of P(gas | ice) / P(gas | steam) will be small

For a word **related** to both ***ice*** and ***steam***, such as **<u>water</u>**, the ratio would be close to one

For words related to **neither** ice nor steam, such as **<u>fashion</u>**, the ratio would be close to one

Probability and Ratio	$k = solid$	$k = gas$	$k = water$	$k = fashion$
$P(k\|ice)$	1.9×10^{-4}	6.6×10^{-5}	3.0×10^{-3}	1.7×10^{-5}
$P(k\|steam)$	2.2×10^{-5}	7.8×10^{-4}	2.2×10^{-3}	1.8×10^{-5}
$P(k\|ice)/P(k\|steam)$	8.9	8.5×10^{-2}	1.36	0.96

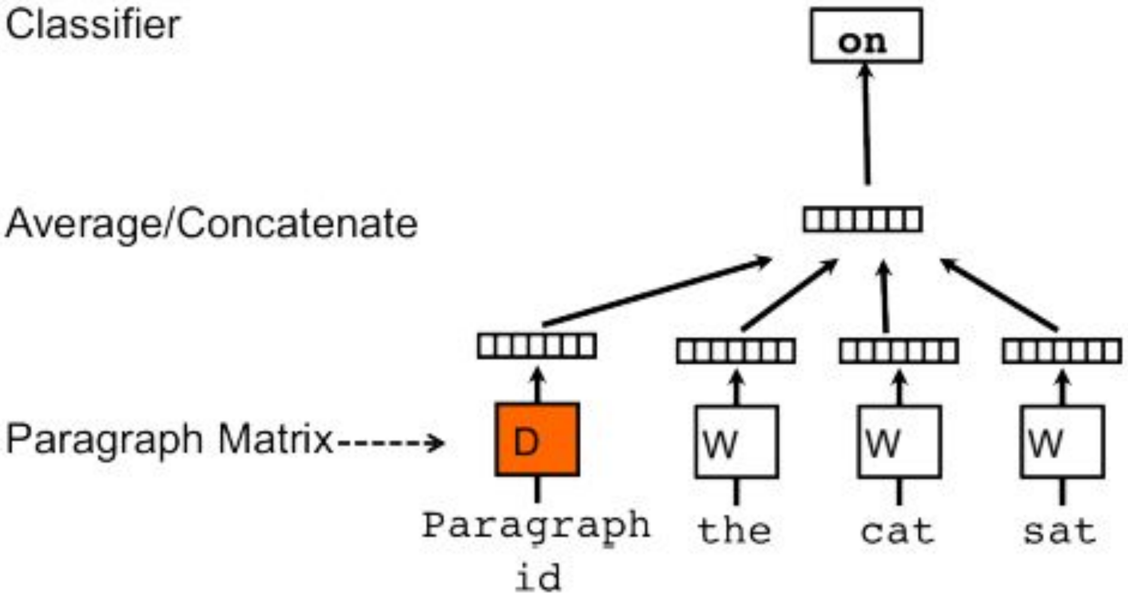
gloVe falls under <u>count based embeddings</u> capturing global co-occurrences and needs an upfront pass of full data during training

source:https://blog.acolyer.org/2016/04/22/glove-global-vectors-for-word-representation/

Word2vec vs gloVe

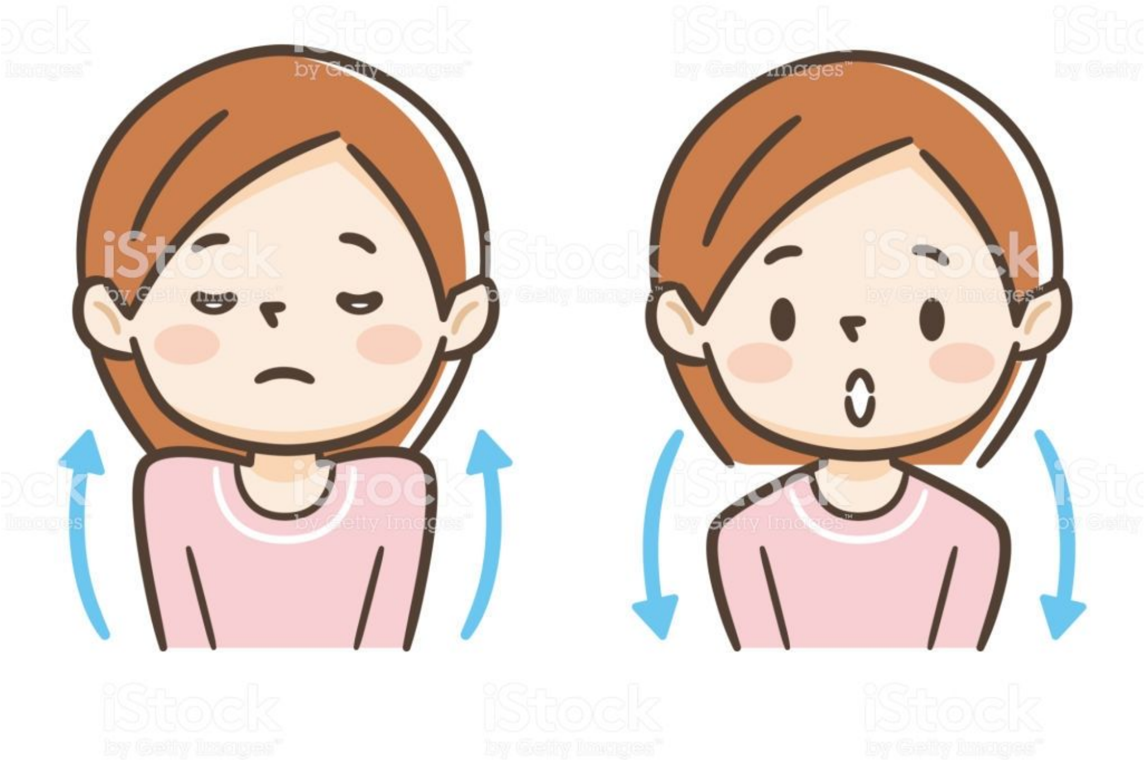
- Word2Vec and GloVe models are very similar in how they work
- Both aim to build a vector space where the position of each word is influenced by its neighboring words based on their context and semantics.
- Word2Vec starts with local individual examples of word co-occurrence pairs
- GloVe starts with global aggregated co-occurrence statistics across all words in the corpus.

One step up - Doc2vec

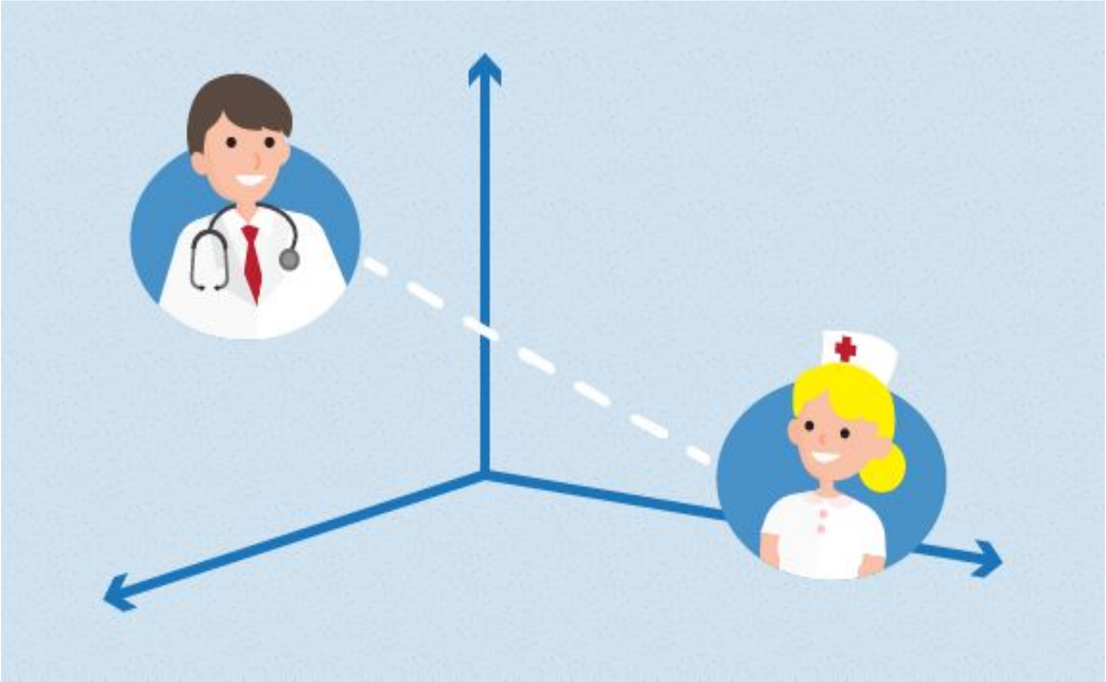


source:https://medium.com/wisio/a-gentle-introduction-to-doc2vec-db3e8c0cce5e

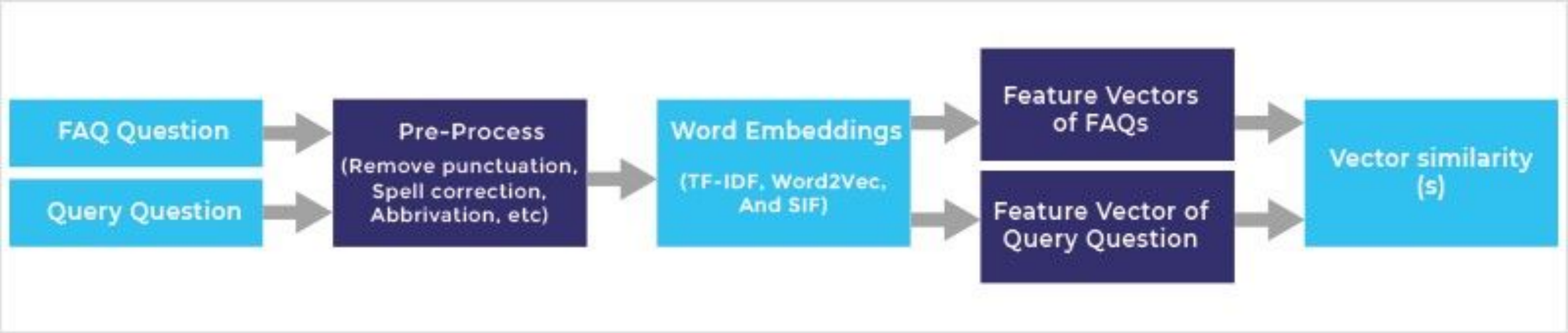
Let’s take a few deep breaths now!



Use-case: Similarity Scoring

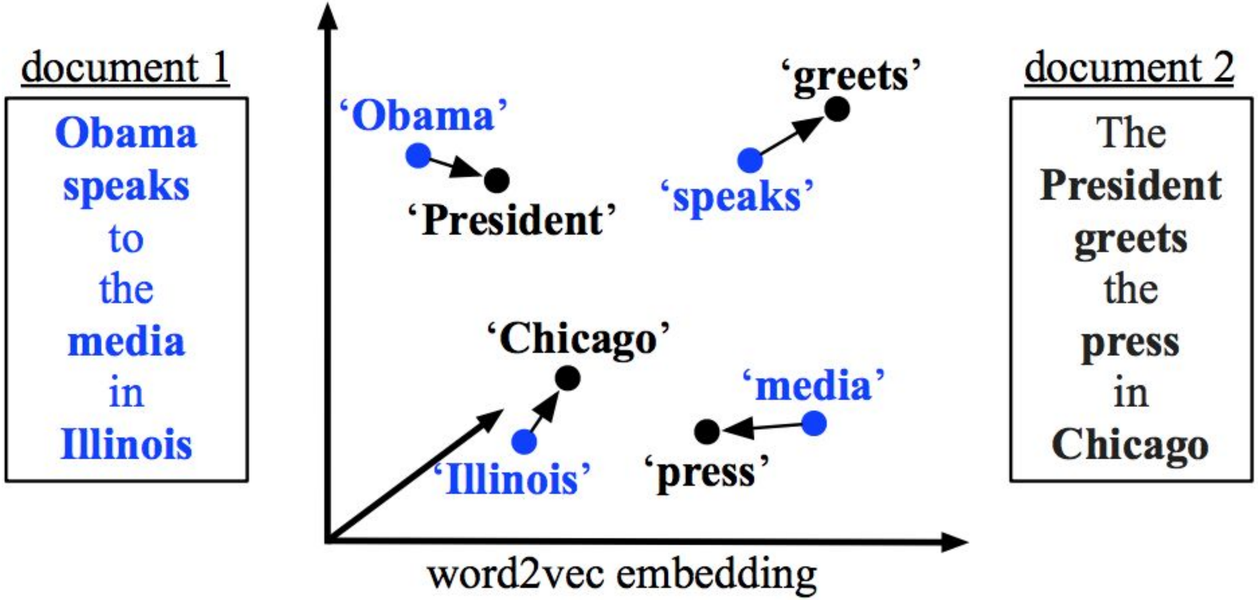


Building Phrase/ Document Similarity Model



Source: https://medium.com/@Intellica.AI/comparison-of-different-word-embeddings-on-text-similarity-a-use-case-in-nlp-e83e08469c1c

Similarity Metrics - Word Mover Distance



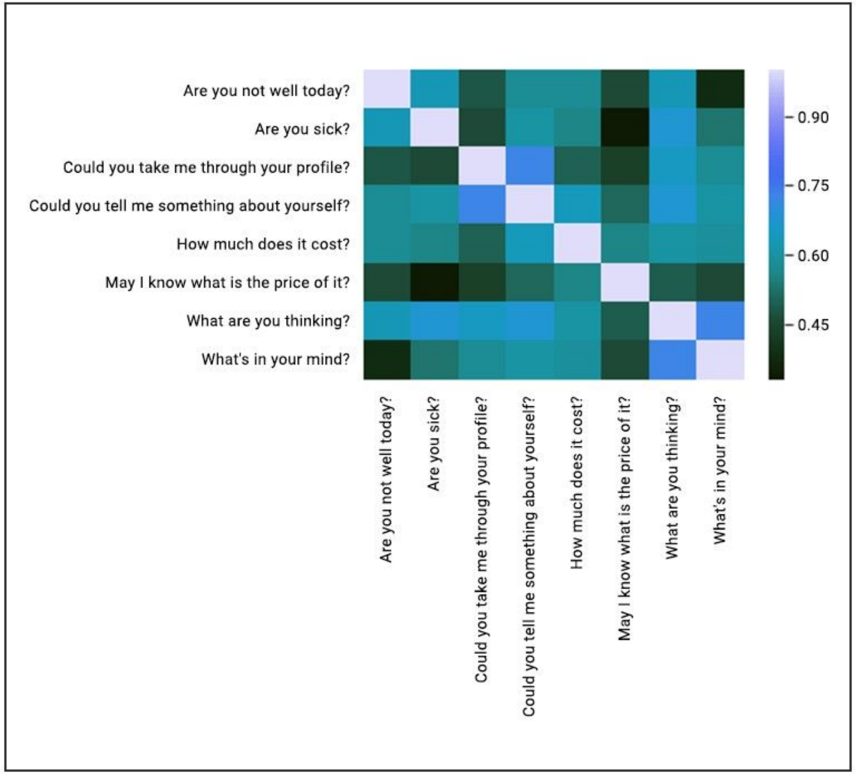
Source: https://medium.com/@adriensieg/text-similarities-da019229c894

Combination of various pre-processing & word-embedding techniques

Pre-Process	Feature Extraction	Similarity Threshold	Questions Answered	Precision
Stop words removal + Punctuation removal + Lemmatization + Spell Correction + Abbreviations	TF-IDF (KB vocab)	0.7	56.85%	90.40%
	word2vec (custom trained)	0.7	27.60%	97.80%
	TF-IDF (KB vocab) + word2vec (custom trained)	0.7	40.12%	96.48%
	SIF	0.7	87.50%	73.72%
	SIF + Clustering	0.8	68.75%	90.62%
	SIF with word2vec or custom trained phrases	0.8	64.11%	93.08%
+ Custom stop words.	SIF + TF-IDF (KB vocab) + word2vec(custom trained)	0.8	74.60%	92.70%

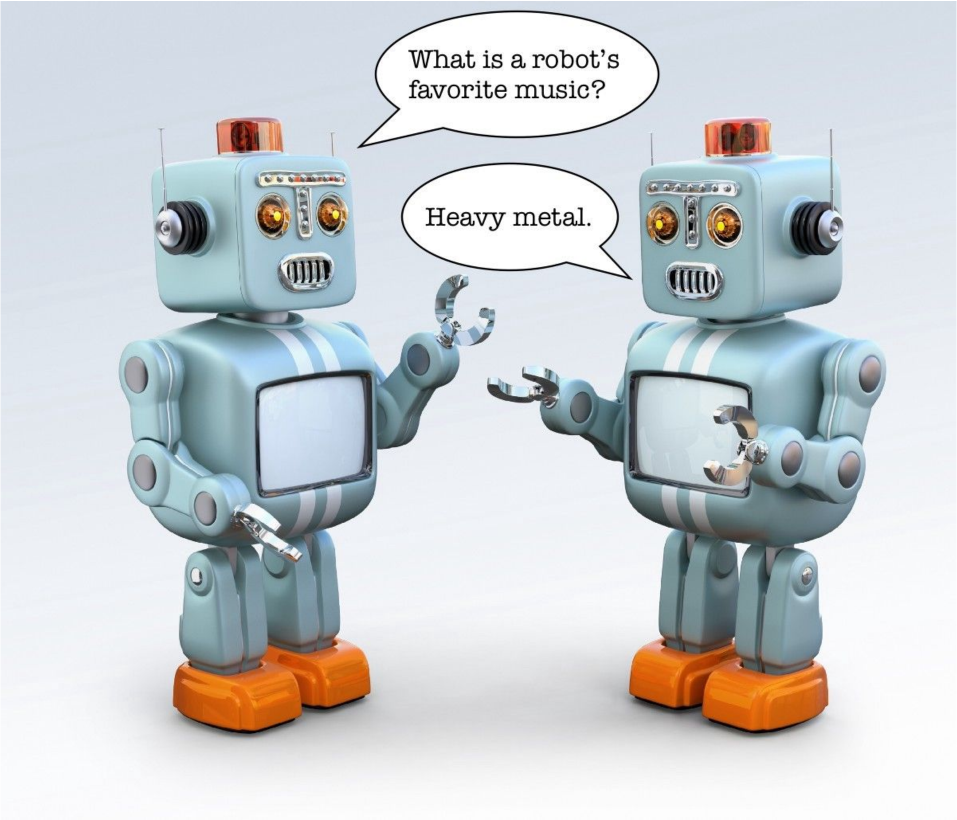
Source: https://medium.com/@Intellica.AI/comparison-of-different-word-embeddings-on-text-similarity-a-use-case-in-nlp-e83e08469c1c

Evaluation - Confusion Matrix

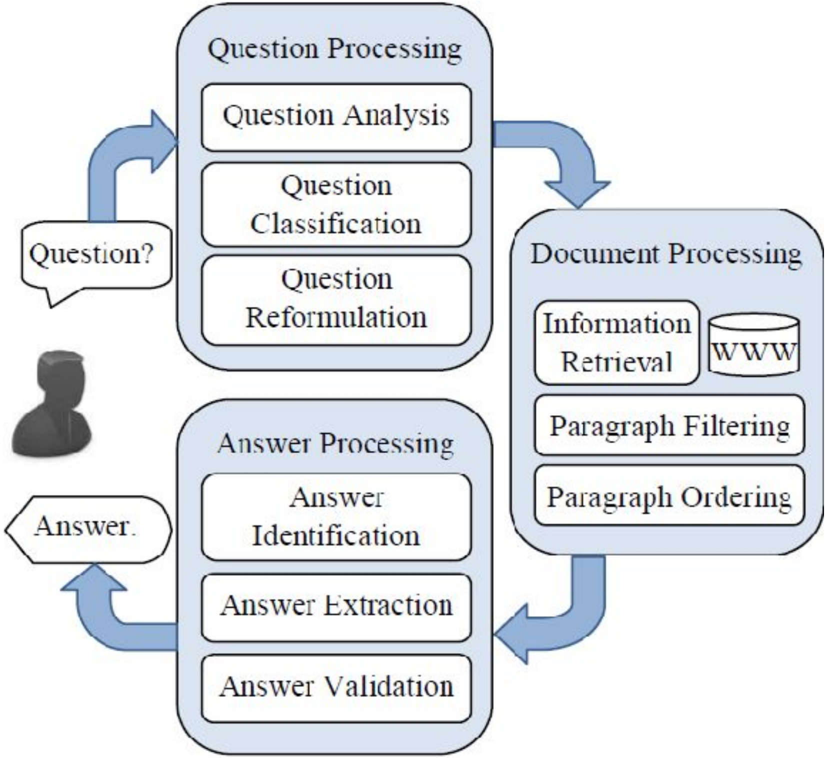


Source: https://medium.com/@Intellica.AI/comparison-of-different-word-embeddings-on-text-similarity-a-use-case-in-nlp-e83e08469c1c

Use-case: Question Answering

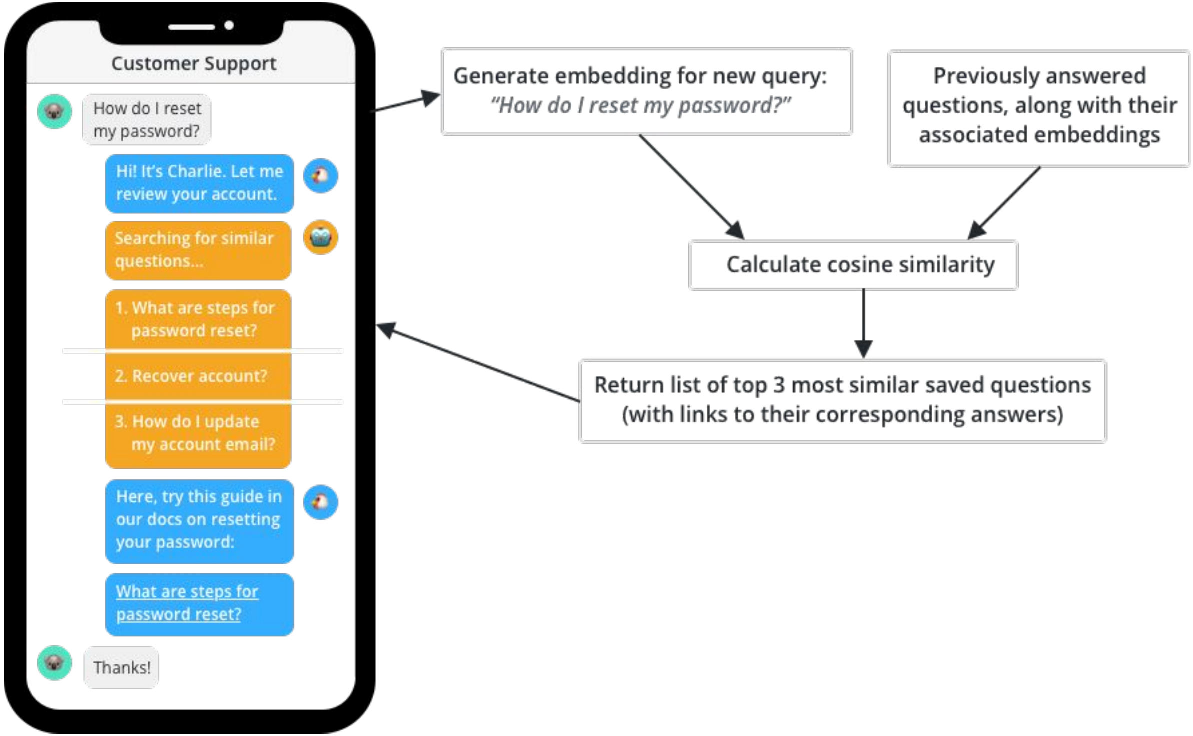


Building QA Model



Source: https://www.semanticscholar.org/paper/Advancements-in-Question-Answering-Systems-Towards-Dhanjal-Sharma/298cf045626c76eef2e408506f5abe8bfb21af17

Applications - Chat Assistant



Source: https://blog.floydhub.com/automate-customer-support-part-two/

Applications - Information Retrieval

Sentence having the right answer

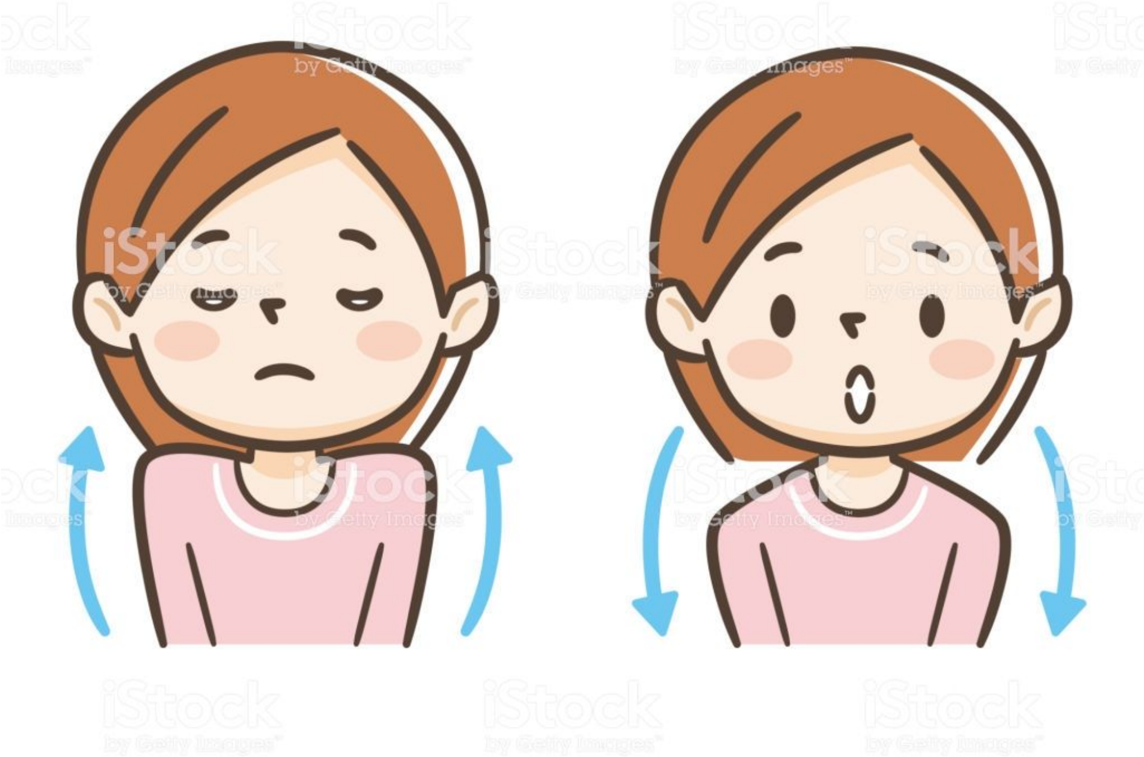
```
'context': 'Beyoncé Giselle Knowles-Carter (/biːˈjɒnseɪ/ bee-YON-say) (bor
n September 4, 1981) is an American singer, songwriter, record producer an
d actress. Born and raised in Houston, Texas, she performed in various sin
ging and dancing competitions as a child,and rose to fame in the late 1990
s as lead singer of R&B girl-group Destiny\'s Child. Managed by her father
, Mathew Knowles, the group became one of the world\'s best-selling girl g
roups of all time. Their hiatus saw the release of Beyoncé\'s debut album,
Dangerously in Love (2003), which established her as a solo artist worldwi
de, earned five Grammy Awards and featured the Billboard Hot 100 number-on
e singles "Crazy in Love" and "Baby Boy".',
'text': 'in the late 1990s’
'question': 'When did Beyonce start becoming popular?'
```

Exact Answer

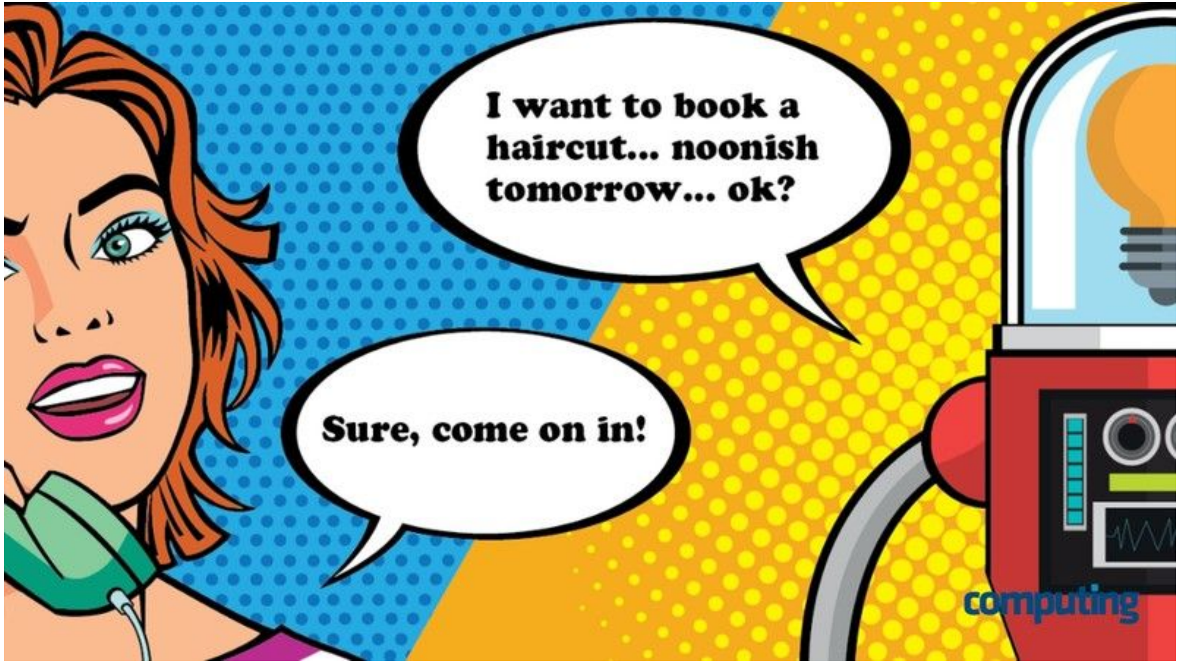
Source: https://towardsdatascience.com/building-a-question-answering-system-part-1-9388aadff507

WOMEN WHO CODE DATA SCIENCE

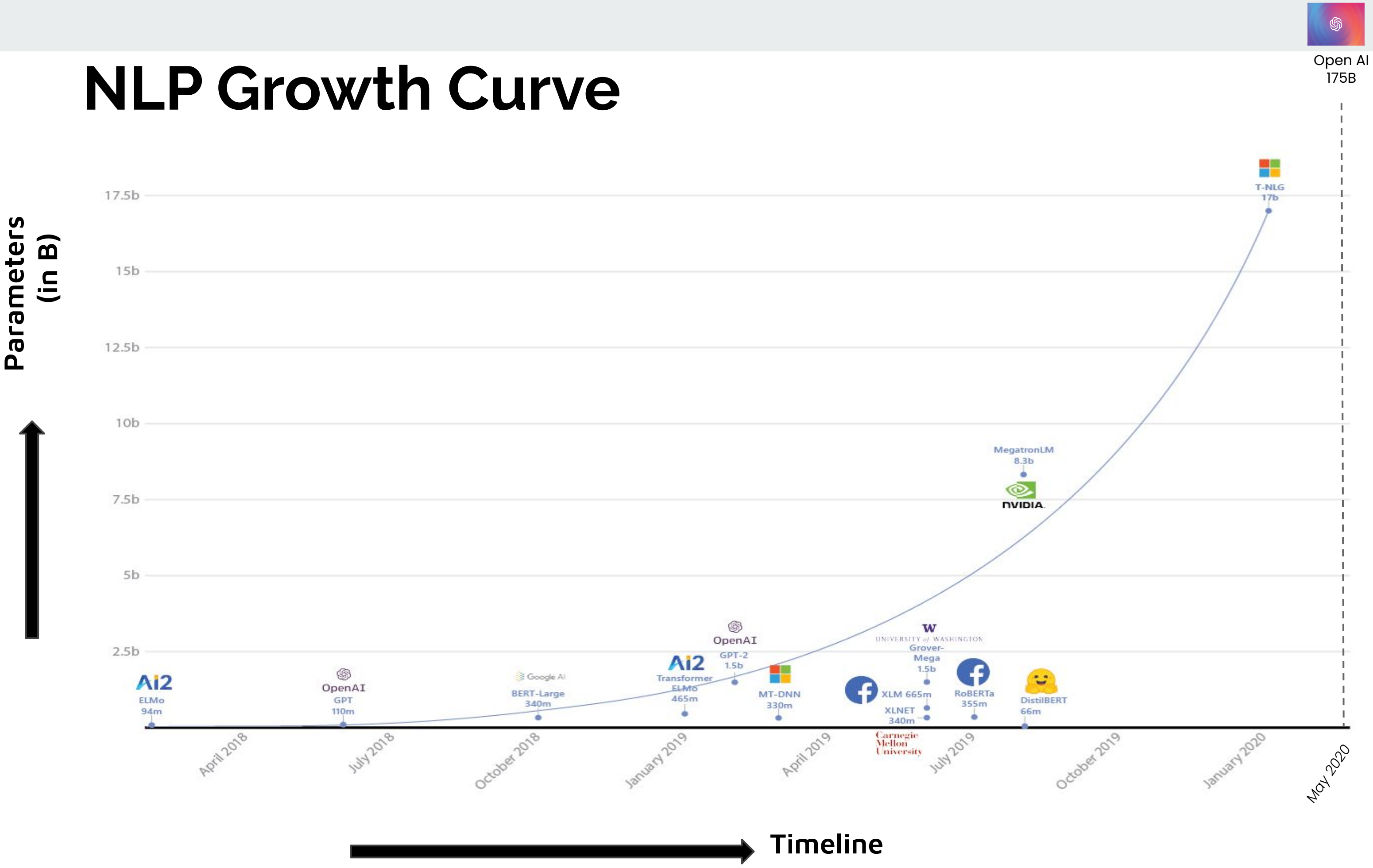
Let’s take a few deep breaths now!



State of NLP today



NLP Growth Curve

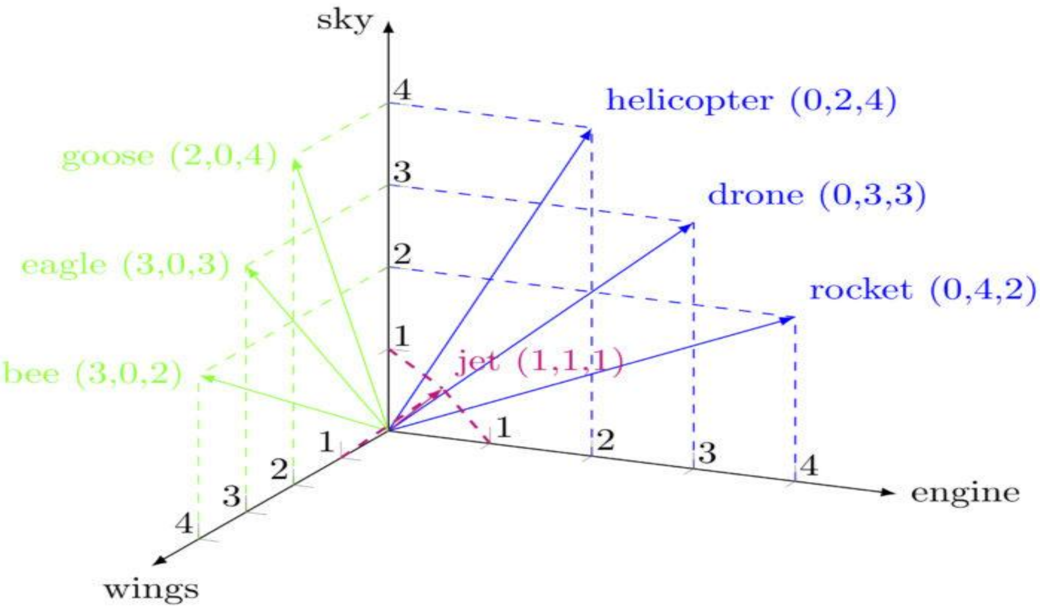


Source: https://www.microsoft.com/en-us/research/blog/turing-nlg-a-17-billion-parameter-language-model-by-microsoft/

WOMEN WHO CODE DATA SCIENCE

Overview of BERT

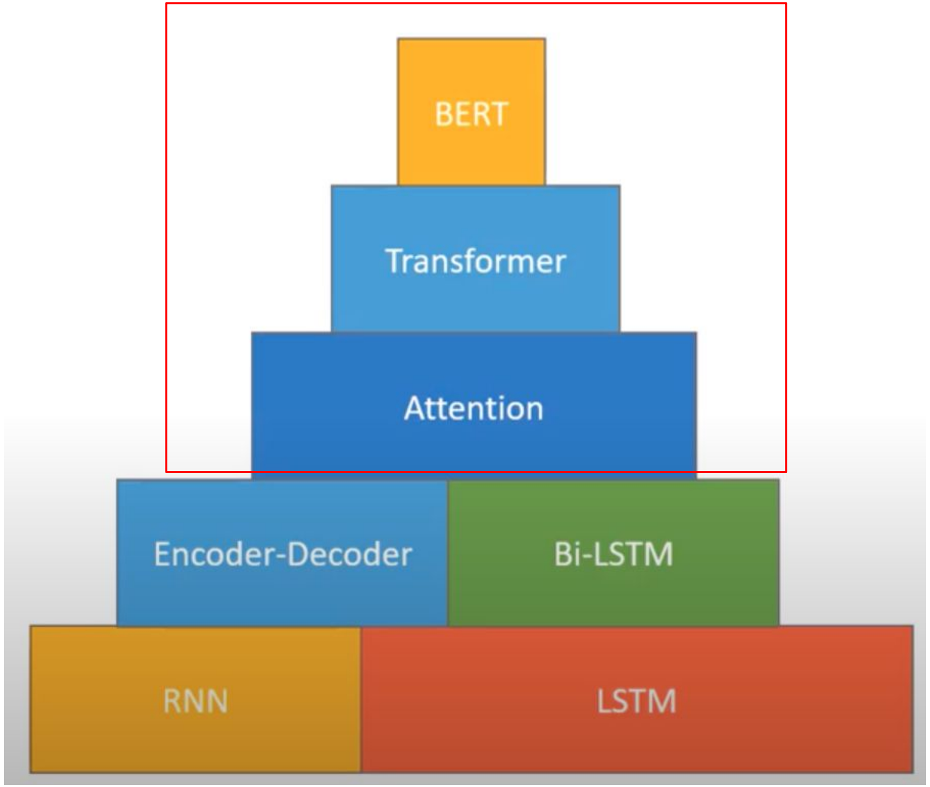
Base Concept



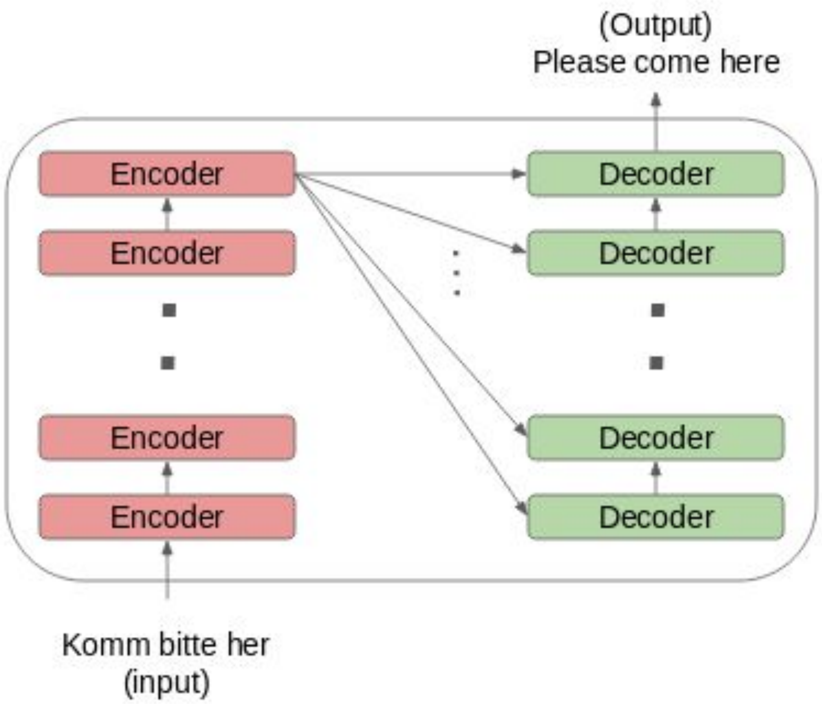
An important concept we learnt - **Word Embeddings** is the base

Word Embedding = Feature Vector representation of a word

BERT Mountain - by Chris McCormick

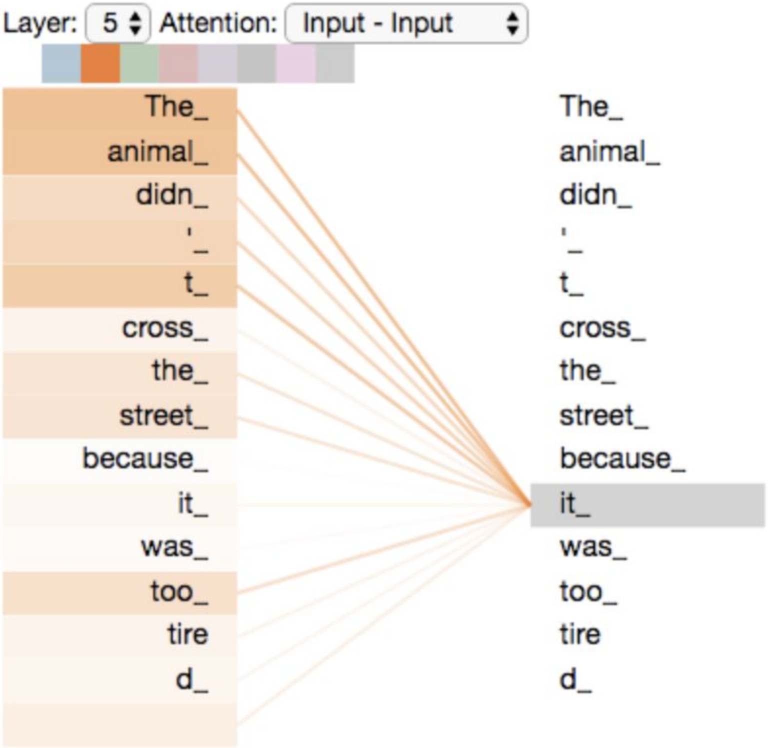


General Idea of Transformers



Source:https://www.analyticsvidhya.com/blog/2019/06/understanding-transformers-nlp-state-of-the-art-models/

Attention is all you need!



Source:http://jalammar.github.io/illustrated-bert/

BERT: Bidirectional Encoder Representations from Transformer

Trained Model:

Wikipedia

(2,500M words),

Book Corpus

(800M words)

Concept:

Bidirectional means that BERT learns information from both the left and the right side of a token's context during training

Architecture:

BERT Base -

-12 layers (transformer blocks),

-768-hidden

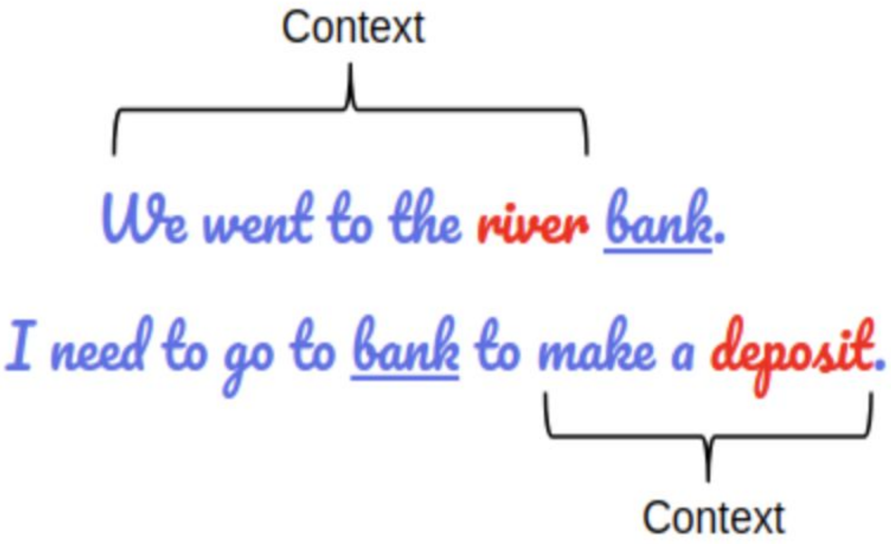
-12 attention heads,

-110M parameters

Trend Setter: First deeply bidirectional, unsupervised, pre-trained on plain text

BERT: Goal

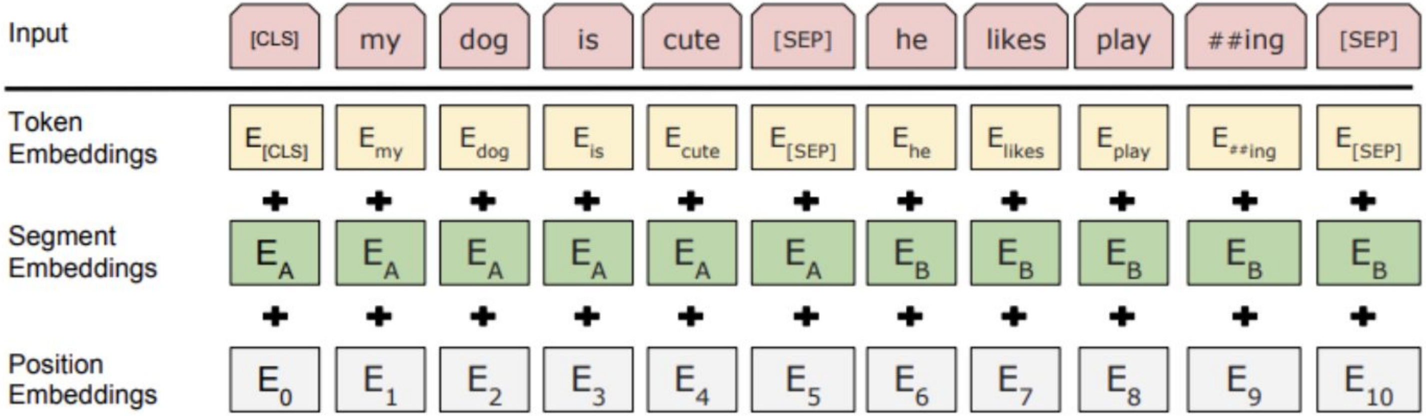
Goal of BERT: Contextual Input Representation and Word Piece Embeddings



BERT: Layers

Three embedding layers

- Token
- Position - word to word relations
- Segment - Sentence to sentence relations



Source:http://jalammar.github.io/illustrated-bert/

BERT: Pre-trained on two NLP tasks

WOMEN WHO CODE DATA SCIENCE

- Masked Language Modeling
- Next Sentence Prediction

MLM: The Strength of Bidirectionality

Input: The man went to the $[MASK]_1$. He bought a $[MASK]_2$ of milk .
Labels: $[MASK]_1$ = store; $[MASK]_2$ = gallon

Next Sentence: model relationships between sentences

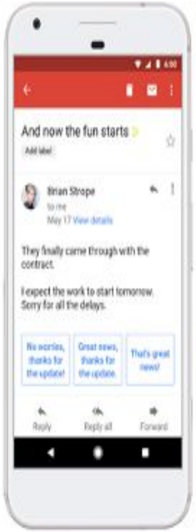
Sentence A = The man went to the store.
Sentence B = He bought a gallon of milk.
Label = IsNextSentence

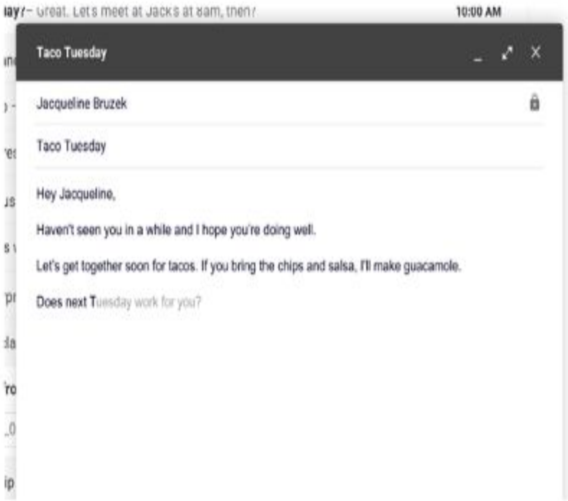
Sentence A = The man went to the store.
Sentence B = Penguins are flightless.
Label = NotNextSentence

Source:http://jalammar.github.io/illustrated-bert/

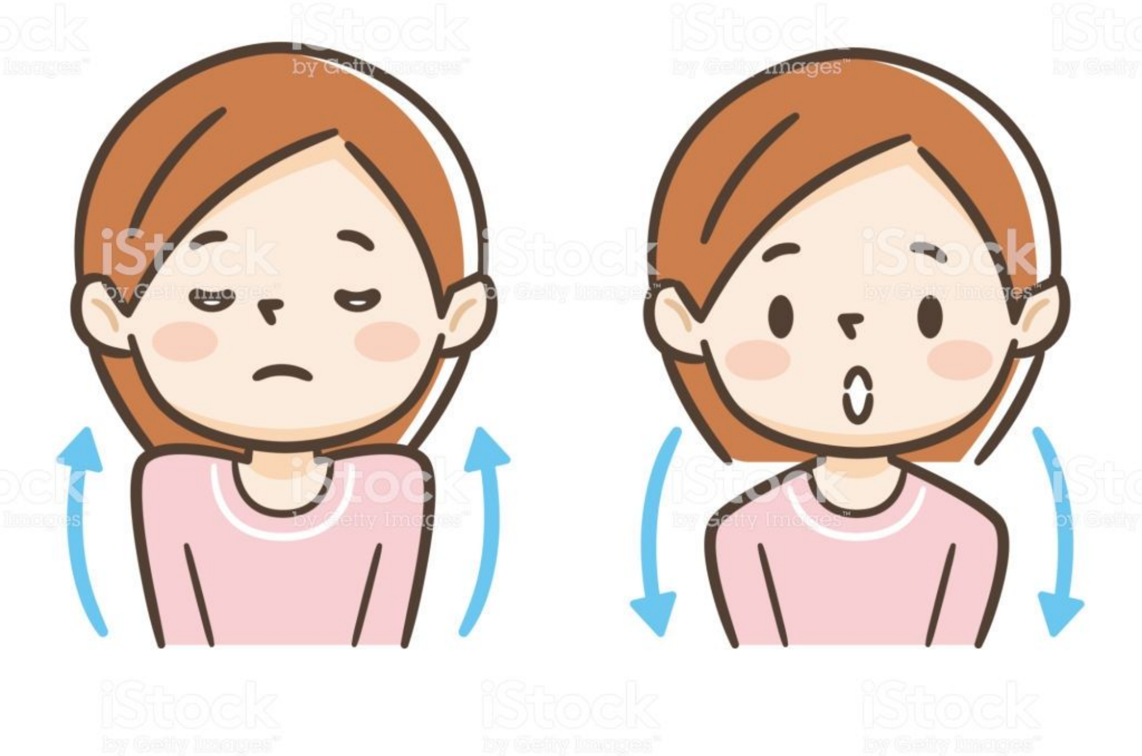
WOMEN WHO CODE DATA SCIENCE

Some important applications





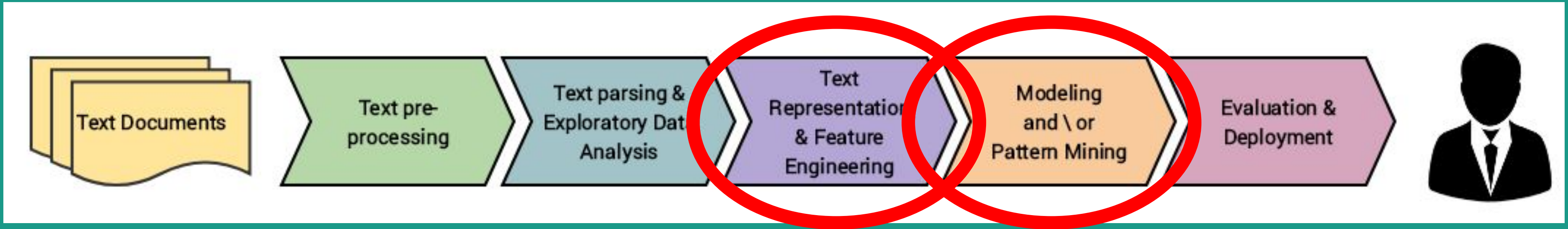
Let's take a few deep breaths now!



WOMEN WHO CODE DATA SCIENCE

4

NLP Workflow



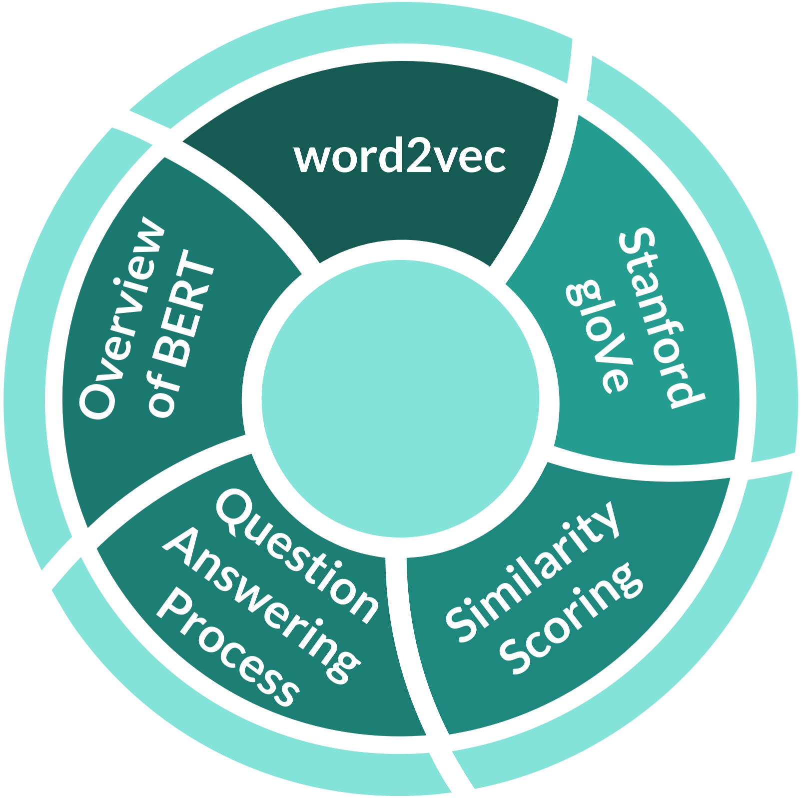
Source: https://towardsdatascience.com/a-practitioners-guide-to-natural-language-processing-part-i-processing-understanding-text-9f4abfd13e72

5

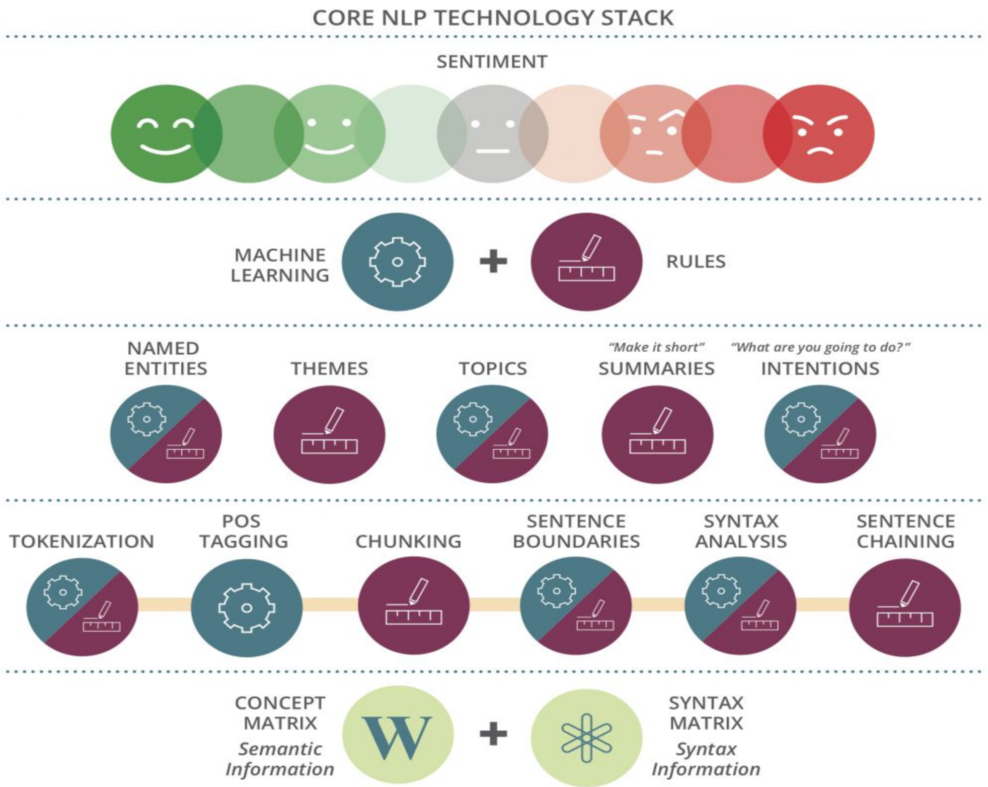
Theory Wrap-up & Next Steps

WOMEN WHO CODE DATA SCIENCE

Recap


word2vec
Stanford gloVe
Similarity Scoring
Question Answering Process
Overview of BERT

Recap



Source: https://www.lexalytics.com/lexablog/machine-learning-natural-language-processing

WOMEN WHO CODE DATA SCIENCE

6

WOMEN WHO CODE DATA SCIENCE

Google Colab Project

https://bit.ly/introtonlp-week4-notebook

WOMEN WHO CODE DATA SCIENCE

Homework #1

Additional Resources

- (Jalammar) The Illustrated Word2vec
- (TowardsDataScience) Intuitive Guide to Understanding Word2vec
- (Chris McCormick) Applying word2vec to Recommenders and Advertising
- (AnalyticsVidhya) Word2Vec using Gensim
- (TowardsDataScience) word2vec-Predict Article Success

WOMEN WHO CODE DATA SCIENCE

See you next week!

Questions?

Join us on Slack and post your questions to the #help-me channel