Icon made by Freepik from www.flaticon.com

自然言語処理
TEACHING MACHINES TO READ

Tomoyuki Jinno
Linkedin

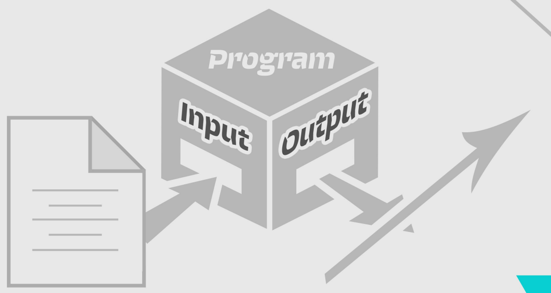
MEVITAE
ROYAL HOLLOWAY UNIVERSITY OF LONDON
SEPnet South East Physics Network

JOB DESCRIPTION VECTOR PROJECT

Aim

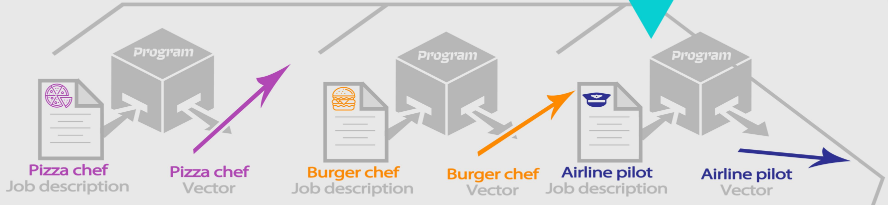
The aim of the project is to wirte a program that takes a job description as an input and output a vector. The direction of the vector correspond to the concepts described in the job description.

Convert document to vector

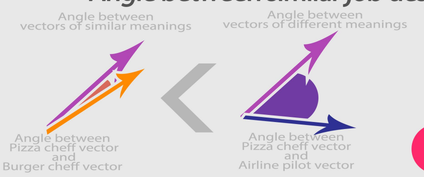


Job description   Vector

Imagine three vectors generated from job descriptions of three different jobs



Angle between similar job descriptions are smaller


Angle between vectors of similar meanings
Angle between vectors of different meanings
Angle between Pizza cheff vector and Burger cheff vector
Angle between Pizza cheff vector and Airline pilot vector

Direction of the job description vectors encode meanings of the docuemnt. Therefore, the direction of vectors generated from documents of similar meaning have similar directions and vectors generated from documents of different meaning have different directions. When two vectors generated from job descriptions of similar jobs. Pizza chef and Burger chef are comped, the vectors have similar directions. However, when vectors generated from job descriptions of vastly different jobs, Pizza chef and Airline pilot are compared, the direction of the vectors are also vastly different, since they contain different concepts.

Direction encodes meaning

Method

1 TF-IDF: Importance of a word for a document

Term freuency - Inverse document frequency

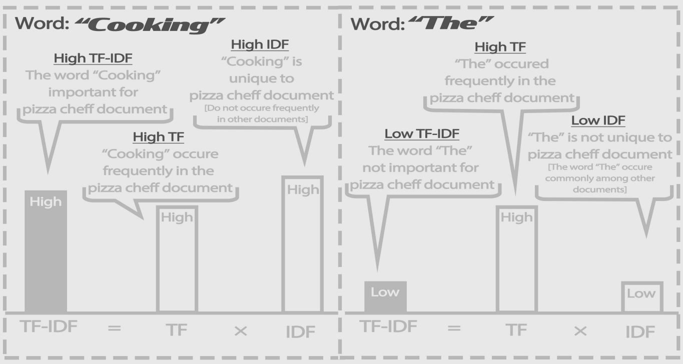
$$TF\text{-}IDF = TF \times IDF$$

Importance of a word for a document

Frequency of the word occurring in a document

Uniqueness of the word: Inverse frequency of a word across a large text corpuse

E.g.



High TF-IDF
The word "Cooking" important for pizza cheff document

High IDF
"Cooking" is unique to pizza cheff document
[Do not occure frequently in other documents]

High TF
"Cooking" occure frequently in the pizza cheff document

High TF
"The" occured frequently in the pizza cheff document

Low TF-IDF
The word "The" not important for pizza cheff document

Low IDF
"The" is not unique to pizza cheff document
[The word "The" occure commonly among other documents]

Document:

Pizza chef
Job description

Figures on left represent how value of the TF-IDF changes as value of TF and IDF change.

Note that value of there is dependent on the word and the document.

Example on left uses "Cooking" and "The" as the word of the TF-IDF and Job description of a Pizza chef as a document.

2 TF-IDF as a Table

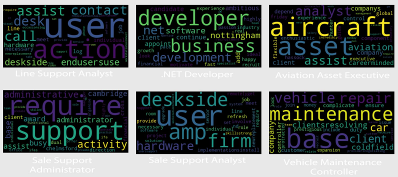
Find TF-IDF values for every words in the English language

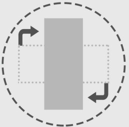
E.g.

Job descriptions	COOKING	THE	··· ··· ···	FLIGHT
	TF-IDF for every words in the English language			
PIZZA CHEF	312	18	··· ··· ···	1
BURGER CHEF	223	2	··· ··· ···	21
AIRLINE PILOT	16	12	··· ··· ···	600

Word clouds on right represent the TF-IDF values for a given job description. The size of the word represents the TF-IDF value of the word for the job description. Greater the size, greater the corresponding TF-IDF value.

Try to guess the job title of the job description used to generate the tf-idf values, by just loking at the word clouds.

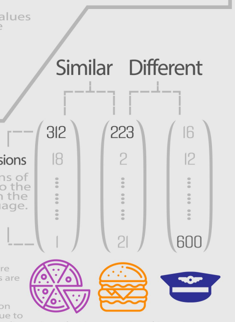




Rotate table to make vectors

To conver the table representation of the TF-IDF values into a vector representation, take transpose of the table representation.

3 TF-IDF as Vectors



Large number of dimensions

The number of dimenisons of the vector is equal to the number of words in the English language.

Curse of dimensionality

Curse of dimensionality is the set of problems that occure when analysing in a high dimensional space. Followings are some of the problems that can be encountered:

Combinatorial explosion, which refer to the phenomenon exponential growth in the complexity of the problem due to exponential increase in possible combinations between dimentions that is encountered in certain problems.

When finding patterns using machine learning, one usualy need to ensure that the data set used to train the algorithm contain at least several data of each combinations of features. Therefore, the amount of data required to train the algorithm increase as the number of features (dimensions) increase.

When measuring Euclidean distance (pythagorean distance in a multi dimensional space) between two points in a high dimensional space, the differences in the distance between different points becomes very small due to the sparse nature of data in a high dimensional space.
(**This is why angles are used instead of distances when comparing job description vectors**)

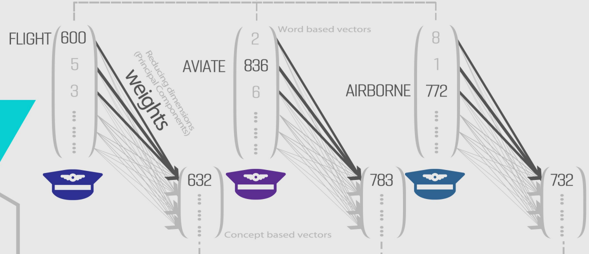
4 Dimension reduction

E.g.

What if synonym is used ?

Imagine TF-IDF vectors generated from three different job descriptions of an airline pilot. If three job descriptions used different words for "flight" , in this case, "aviate" and "airborne", The generated TF-IDF vectors look largely different from each other, even though concepts described in the three job descriptions should be similar. This issue can be resolved by reducing the number of dimensions.

TF-IDF vectors of similar jobs are different



Reduced vectors are similar

How dimensions are reduced

Singular value decomposition

SVD is used to find principle components. **Principal components** can be thought of as linear transformations that **maximise the variance of data**, hence the Neural Network like diagram. The PC can be found in descending order from the PC that encode the largest variance to the least. Therefore, to reduce dimensions, first few PCs can be used to represent majority of the variace in data.

How does it combine dimensions by similar meaning?

Distributional Hypothesis states that words occuring in similar context have similar meanings..When you find a word that you don't know the meaning of, you can often deduce the meaning by finding other words that you know of that can be used in the same context. This is due to the distributional hypothesis.

SVD maximises the variance by combining ombining co-occuring features. When SVD is carried out on TF-IDF vectors, TF-IDF values of highly co-occuring words are combined. Assuming that words with high co-occurance can be used interchangeably, TF-IDF values of similar words are combined into same dimension.
The method of applying SVD on TF-IDF vectors is called **Latent Semantic Analysis**.

5 Concept based vectors