

Application of deep neural networks
to natural language processing

Sentiment analysis

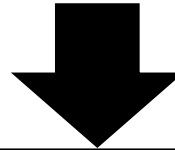
گروه دایچه . dayche.com



Natural language processing



- There are many website that ask us to give reviews or feedback about there product when we are using them. like:- Amazon, IMDB.
- We also use to search at google with couple of words and get result related to it.
- There are some sites that put tags on the blog related the material in the blog.




- These are some example of text processing.
- Text processing – we use text processing to do sentiment analysis, clustering similar words, document classification and tagging.
- Natural language processing (NLP) is a general term dealing with human communication tools like speech, image, text, signs and anything that contain textual information. Therefore, it is something beyond text processing which only deal with textual documents.

تولید محتوا: وحید محمدزاده ایوقی

daychegroup 

daychegroup 

گروه دایکه | dayche.com 

How computers perceive textual content




- As we read any newspaper we can say that what is the news about but how computer will do these things?
- Computers can match string to find whether some words are similar. How do they do that? Using encoding procedure.
 - This mechanism is arbitrary and provide no useful information. For example the word cat may be perceived by computer using ID143 and the word dog by ID144. However, how the relationship between them will be understood? Both are animal after all.
 - So it would not be sufficient.
- Like images and speech signals, an encoding mechanism should be exist.
 - Images are encoded based on integer values corresponding to color each pixel represent (RGB).
 - For text data, we need word embedding to do the same task.

تولید محتوا: وحید محمدزاده ایوقی

daychegroup 

daychegroup 

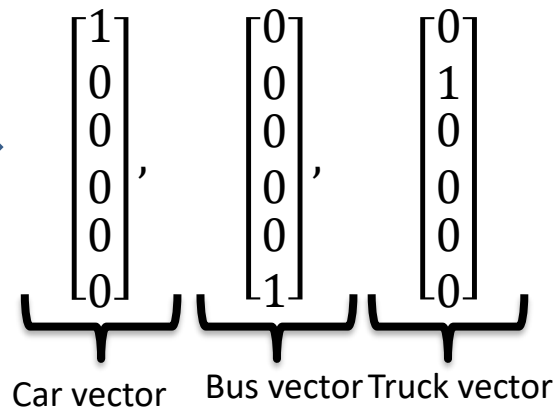
گروه دایکه | dayche.com 

Numerical representation



- We need a process to be able to state the similarity between different words in a given sentence.
 - String matching will not reflect this concept, since the word man and the word boy is not alphabetically equivalent, although they are similar.
 - One hot encoding - represent each words with a binary vector
 - Create a distinct bag of words in order. (apparently if the size of corpus is large, then the word representation will be large and sparse.)

Car
Truck
Jeep
Walk
Bike
Bus



- What about new words which are not in training set of corpus
- Information loss is 100%

تولید محتوا: وحید محمدزاده ایوقی

daychegroup

daychegroup

dayche.com | گروه دایچه

One hot encoding – problem


- Other than the mentioned problem with one hot encoding, the most important problem is the orthogonality of representation.
 - The dot product of obtained representation is zero, meaning there is no relationship between different words. This is not basically true since the car and jeep is similar to each other.




- The need for word representation in such a way that each vectors preserve its similarity to other words, if exist.
- Word embedding is a name for this process. A process which yield to a numerical representation of words in such a way that similar words have relevant representation

تولید محتوا: وحید محمدزاده ایوقی

daychegroup 

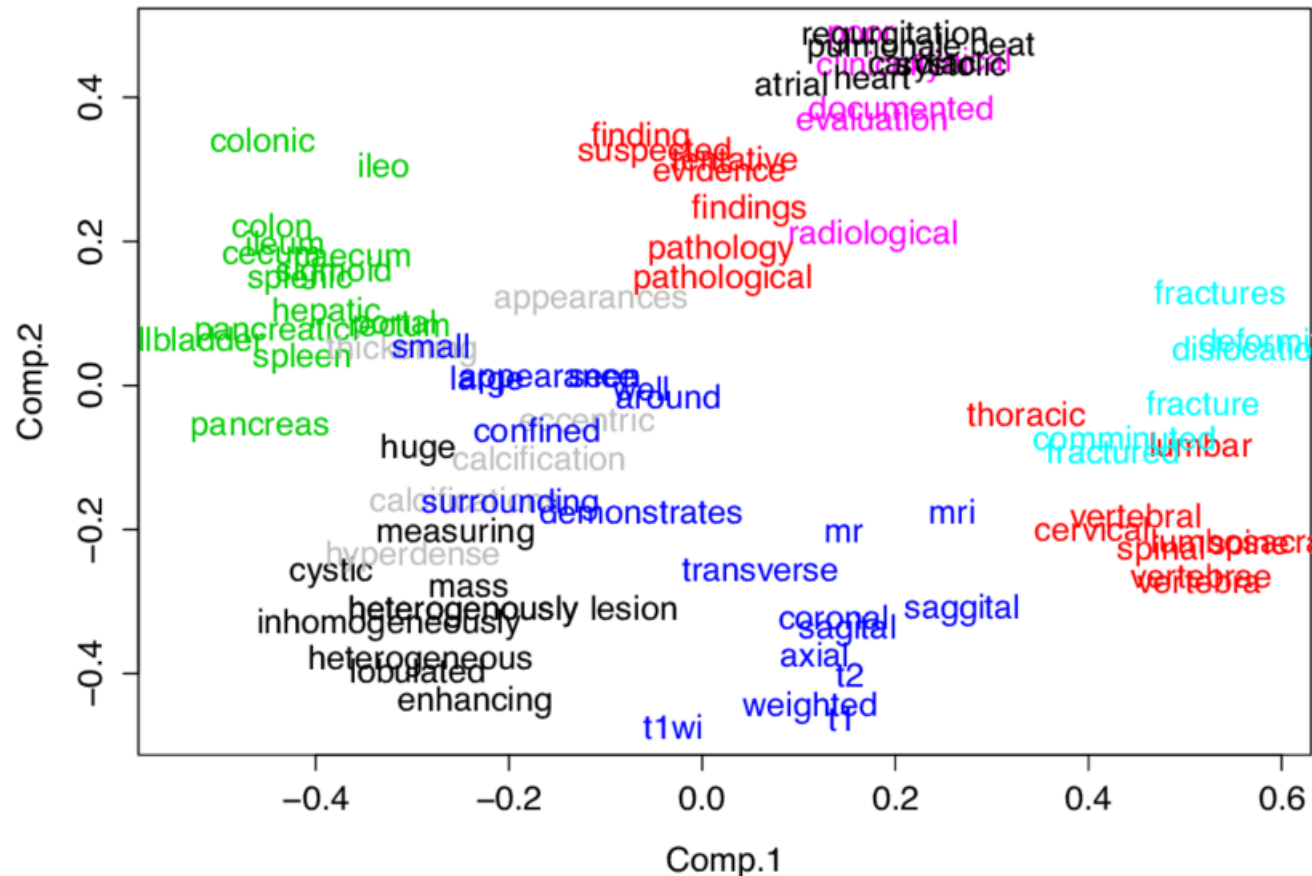
daychegroup 

dayche.com | گروه دایکه 

Word embedding



- Word embedding – a numerical representation of data



A numerical representation which preserve contextual relationship between different words.

تولید محتوا: وحید محمدزاده ایوقی

daychegroup

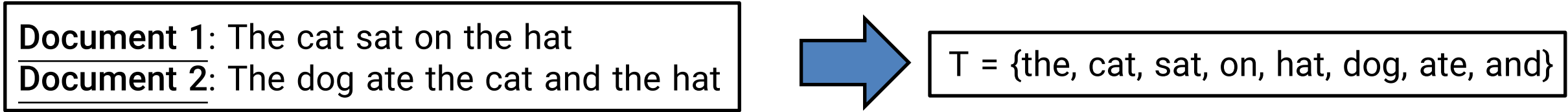
daychegroup

گروه دایچه | dayche.com

Frequency based embedding



- Count vector– a count vector learns vocabulary using their occurrence frequencies in a given set of documents.
- Consider we have D documents and T is the number of different words in our vocabulary sets, our data matrix will have D rows and T columns calling tokens.



- So the obtained data matrix is of 2 by 8 dimension.

	The	cat	sat	on	hat	dog	ate	and
D1	2	1	1	1	1	0	0	0
D2	3	1	0	0	1	1	1	1

Word vector

TF-IDF vectorization

- TF-IDF standing for Term Frequency – Inverse Document Frequency
 - A method for quantifying word frequency based on their significant – a way for reweighting tokens
 - Count vectors suffer from dominating effect of stop words in a given set of documents

$$\text{score} = t_f \times \log \frac{N}{n}$$




t_f , N and n stand for the number of occurrence of term t in each document, number of documents, and the number of documents contain term t respectively.

	The	cat	sat	on	hat	dog	ate	and
D1	0	0	0.05	0.05	0	0	0	0
D2	0	0	0	0	0	0.03	0.03	0.03

تولید محتوا: وحید محمدزاده ایوقی

daychegroup 

daychegroup 

گروه دایچه | dayche.com 

Word co-occurrence matrix



- Words co-occurrence matrix describes how words occur together that in turn captures the relationships between words.
 - Word co-occurrence matrix is not a word representation itself, yet it can result in an appropriate word vector.
 - It is also called bigram frequency.
 - For a corpus of size N, the word occurrence matrix would of N by N matrix. Elements of this matrix are computed based on conditional frequency $P(w_{next}|w_{current})$.

Document 1: The cat sat on the hat

Document 2: The dog ate the cat and the hat

	The	cat	sat	on	hat	dog	ate	and
The	0	2	0	1	2	1	1	1
Cat	2	0	1	0	0	0	0	1

تولید محتوا: وحید محمدزاده ایوقی

daychegroup

daychegroup

گروه دایکه | dayche.com

Bigram frequency – word representation




- As it has already mentioned, the bigram frequency matrix is not word embedding vector itself. However it can be used for representing word vector.
- Upon creating bigram frequency, perform PCA, or rather any type of dimension reduction. The result would be a word representation of dimension k.

	x_1	x_2	x_3	x_4
The	0.015252	0.065987	0.0659451	0.98542
Cat	0.32652	0.1548	0.2365	0.00856
Sat	0.032465	0.0006598	0.065841	0.45213
On	0.032656	0.000358	0.036598	0.06545
Hat	0.8587	0.02537	0.45825	0.03658
Dog	0.09584	0.025658	0.036659	0.033658
Ate	0.025488	0.24158	0.033326	0.75548
And	0.06598	0.7546	0.89595	0.123546

تولید محتوا: وحید محمدزاده ایوقی

daychegroup 


daychegroup 

dayche.com | گروه دایکه 

Continues bag of words (CBOW)



- In case, the vocabulary size is large, the word co-occurrence matrix would be large.
 - Storage problem
- Represent a word vector based on the context, or rather its surrounding words. CBOW obtain the word embedding based on the prediction of current word using the context.

It is a pleasant day  Find an appropriate representation for the word pleasant by which one be able to predict the word day., or rather context

Pairs of context and target = $\{([it, is], a), ([is, a], pleasant), ([a, day], pleasant), \dots\} \sim \underline{(\text{context words, target})}$


Context size – a hyper parameter

- CBOW model accept context words as input and make an attempt to predict the target.
- Input and output vector are one-hot-encoded vector.
- The word representation vector is the output of hidden layer of model.

تولید محتوا: وحید محمدزاده ایوقی

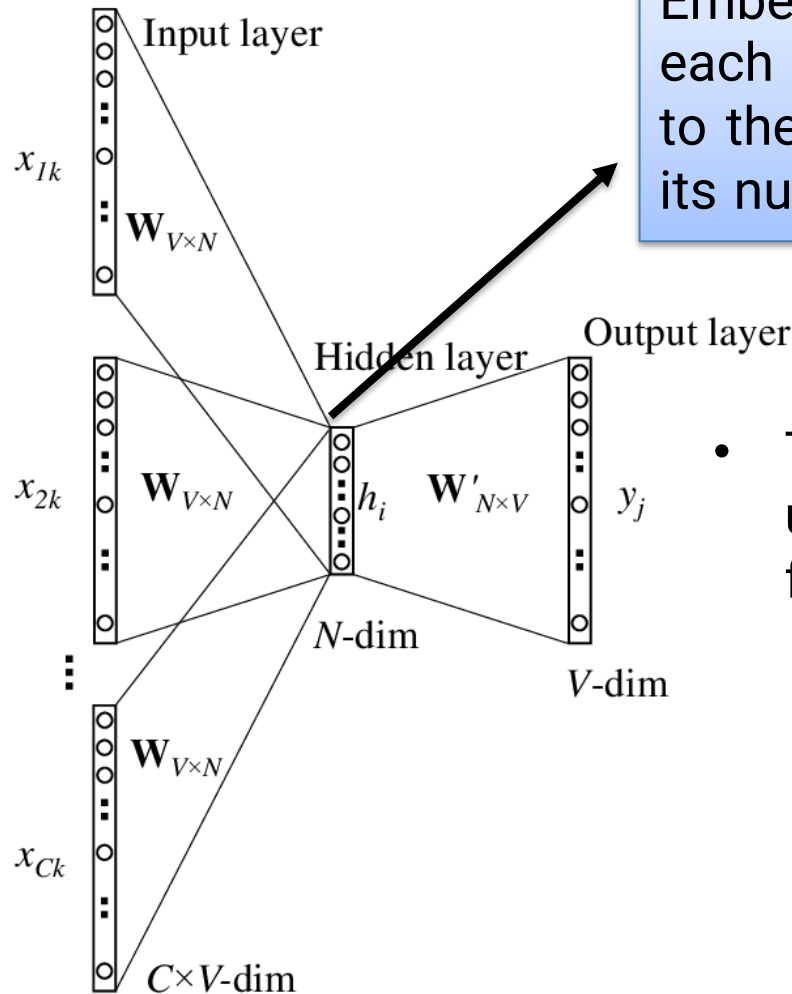
daychegroup 

daychegroup 

گروه دایچه | dayche.com 

CBOW structure

- Context words – there are C context words associated with a single current word
 - The parameter C is a hyper parameter
- According to the given scheme, for each one hot encoded word, there is a representation, meaning we will have C representation on bottleneck. How this would be merged?
- Taking an average in train time.



Embedding layer – after training each word can be given as input to the network. Network will return its numerical representation.

- The network is trained using cross entropy loss function

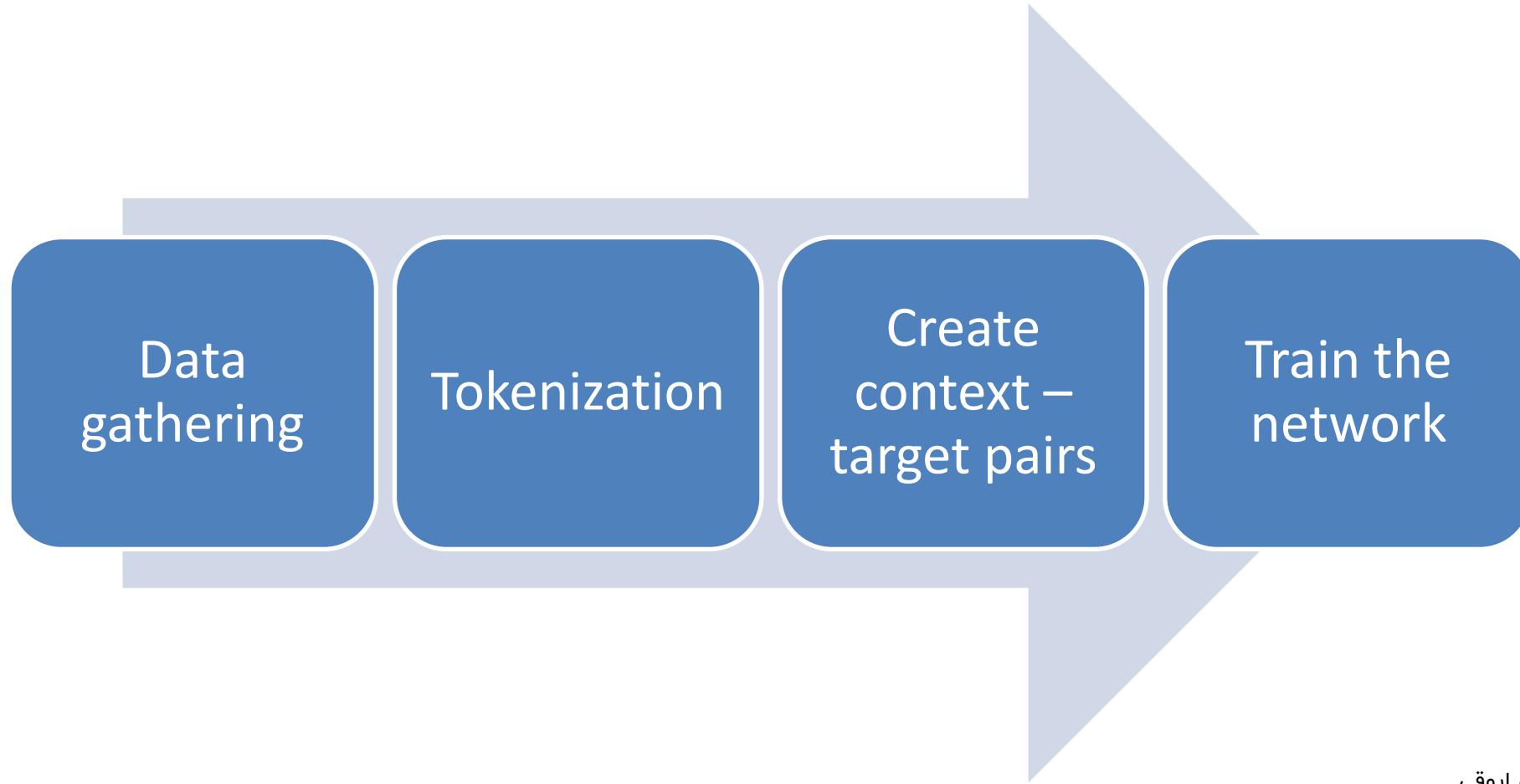
تولید محتوا: وحید محمدزاده ایوقی

daychegroup

daychegroup

گروه دایچه | dayche.com


CBOW pipeline – Word embedding



تولید محتوا: وحید محمدزاده ایوقی

daychegroup 

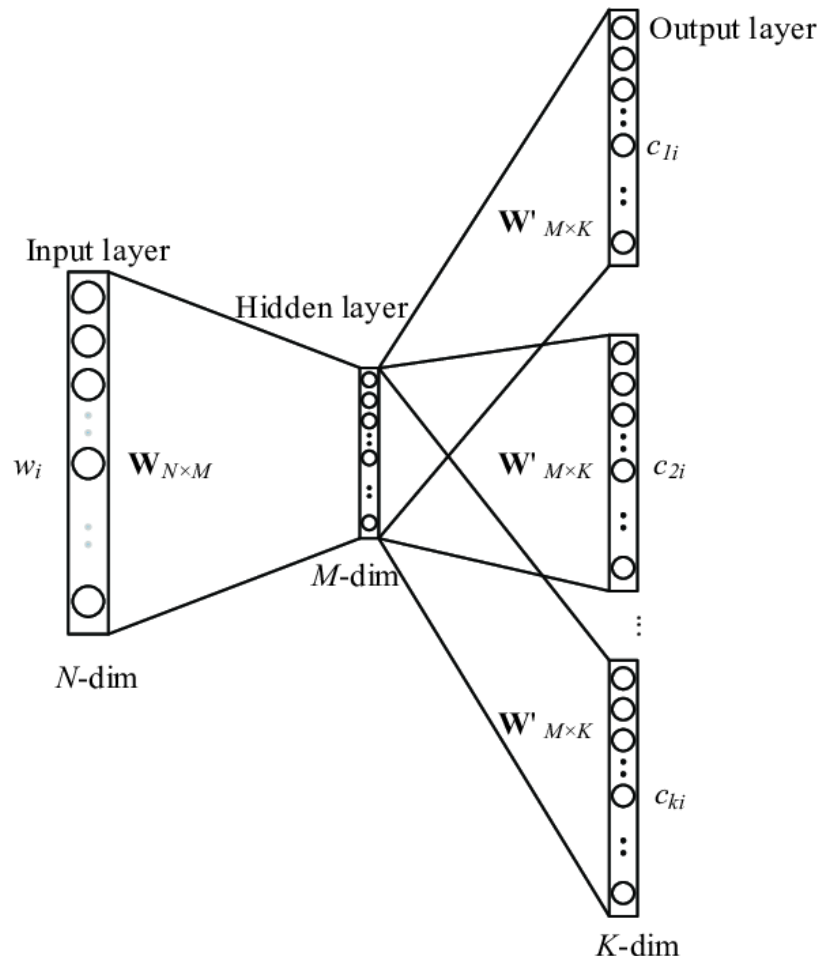
daychegroup 

گروه دایکه | dayche.com 

Skip-gram



- Another way to construct a word embedding model is the use of skip gram.
- Skip gram reverses the mechanism of CBOW.




- The training mechanism of skip-gram is the same as CBOW.
- Can you find a similarity between these two different structures and other machine learning model?

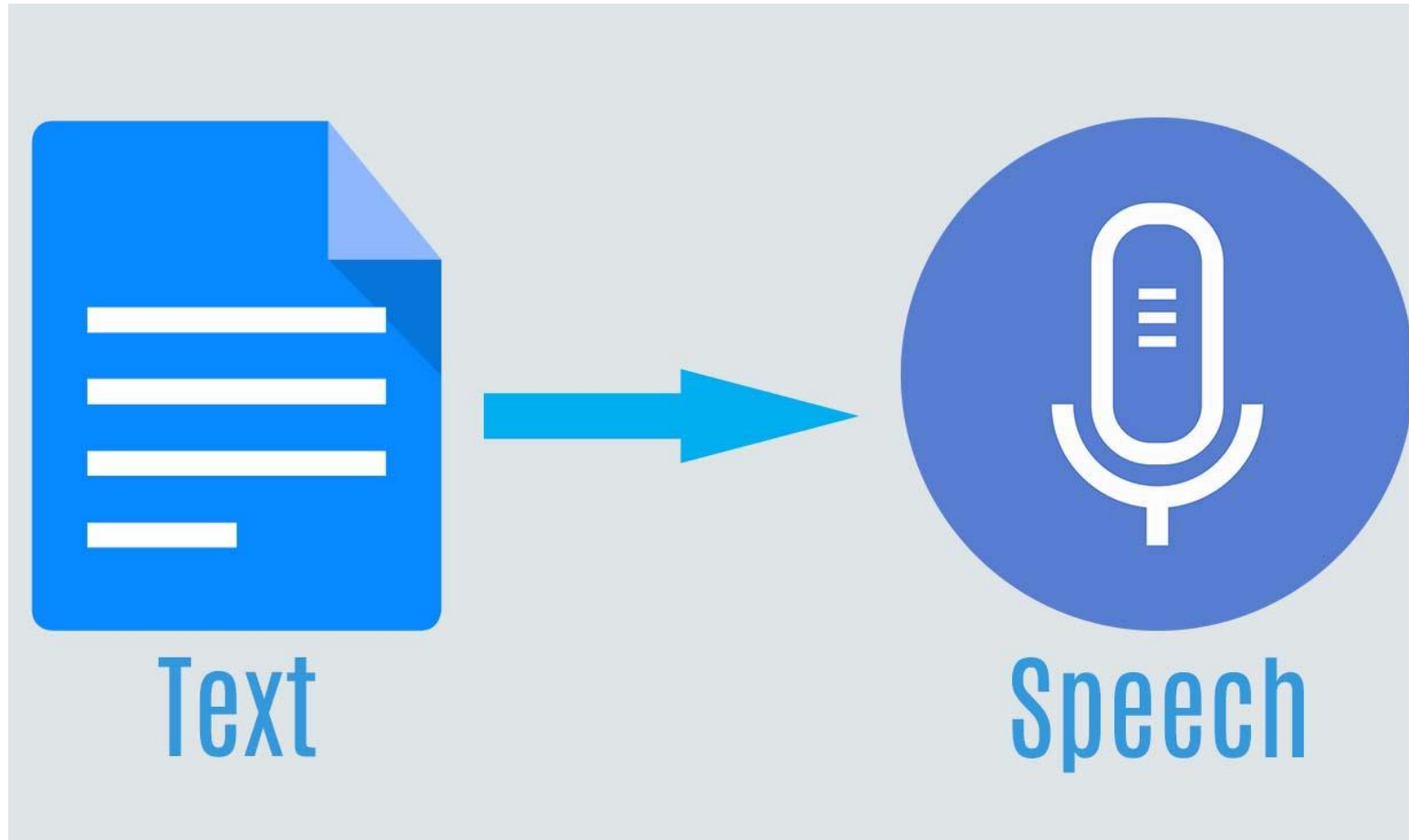
تولید محتوا: وحید محمدزاده ایوقی

daychegroup 

daychegroup 

dayche.com | گروه دایچه 


Application – Text to speech



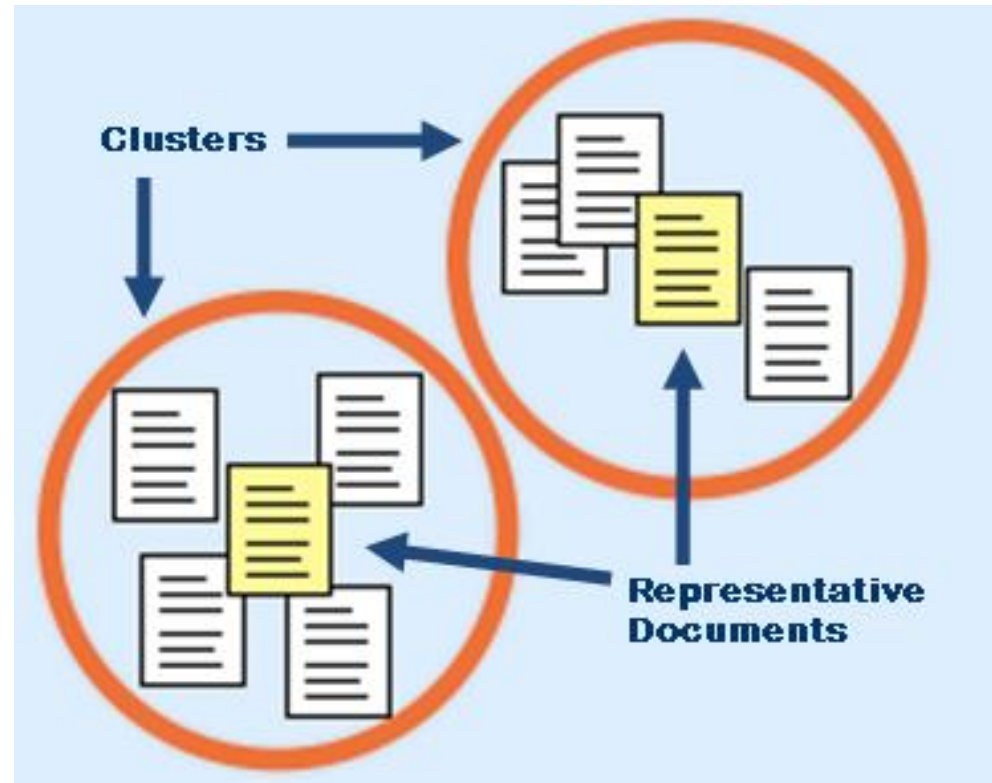
تولید محتوا: وحید محمدزاده ایوقی

daychegroup 

daychegroup 

گروه دایکه | dayche.com 

Application – Document clustering



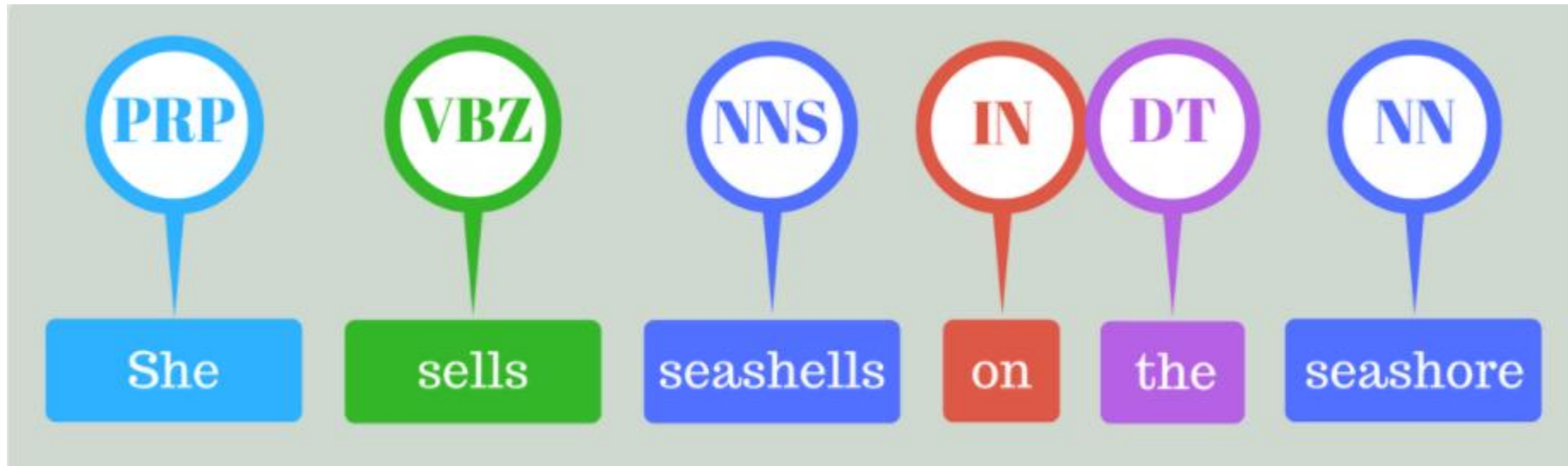
تولید محتوا: وحید محمدزاده ایوقی

daychegroup

daychegroup

گروه دایکه | dayche.com


Application – tagging



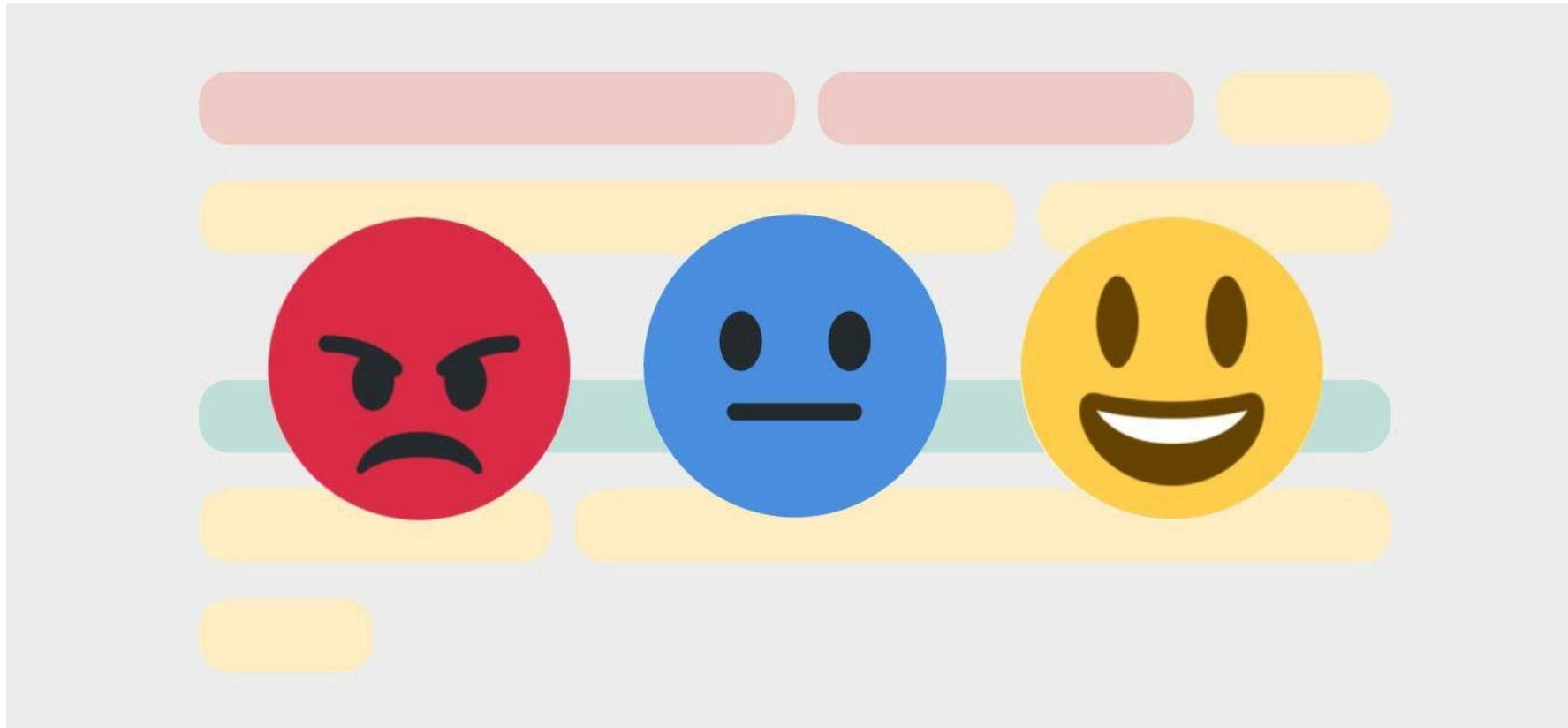
تولید محتوا: وحید محمدزاده ایوقی

daychegroup 

daychegroup 

dayche.com | گروه دایکه 

Application – sentiment analysis



تولید محتوا: وحید محمدزاده ایوقی

daychegroup 

daychegroup 

dayche.com | گروه دایکه 