# Lab 1 - Cosine Similarity & Accuracy: a Focus on the Analogy Task

Alberto Testoni, 9th November 2020

We want to find the nearest neighbours of a word in a vector space. What we need:

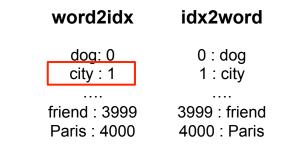
- 1. A matrix of all the word embeddings
- 2. A "dictionary" that maps each word to a row in the matrix, and vice versa
- 3. A distance function (cosine similarity)

	Length of the word embeddings						
Vocabulary size (# words)	0.1	-0.3	0.2		0.1	0.6	0.8
	0.2	0.4	0.1		0.2	0.5	0.3
	-0.5	-0.8	0.4		-0.8	0.4	0.5
Voca	0.8	0.3	0.2		0.1	0.4	-0.9

word2idx	idx2word
dog: 0 city:1	0 : dog 1 : city
friend : 3999 Paris : 4000	 3999 : friend 4000 : Paris

What is the word embedding of "city"?

0.1	-0.3	0.2	 0.1	0.6	0.8
0.2	0.4	0.1	 0.2	0.5	0.3
-0.5	-0.8	0.4	 -0.8	0.4	0.5
0.8	0.3	0.2	 0.1	0.4	-0.9



## Which word corresponds to the last row in the matrix?

0.1	-0.3	0.2	 0.1	0.6	0.8
0.2	0.4	0.1	 0.2	0.5	0.3
-0.5	-0.8	0.4	 -0.8	0.4	0.5
0.8	0.3	0.2	 0.1	0.4	-0.9

word2idx	idx2word		
dog: 0 city : 1	0 : dog 1 : city		
friend : 3999 Paris : 4000	 3999 : friend 4000 : Paris		

#### Let's Look at the Code!

How do we compute the nearest neighbours of a word in a vector space?

https://colab.research.google.com/drive/1y9PtwOZ2E2k5aThj5cmVFPIDD24ZT-NI?usp=sharing

## The Analogy Task

• A proportional analogy holds between two word pairs:

```
x : y = a : b (x is to y as a is to b)
```

• For example:

```
man : king = woman : X
```

• An interesting property of word embeddings is that analogies can often be solved simply by adding/subtracting word embeddings.

#### Let's Look at the Code!

How do we solve an analogy with word embeddings?

## Analogy Test Set (Mikolov et al., 2013)

- We will use the same dataset as in Baroni et al., 2014: <u>http://www.fit.vutbr.cz/~imikolov/rnnlm/word-test.v1.txt</u> (open the file and search for ":" to have a look at all the analogy types)
- We will evaluate the word embeddings using the accuracy metric:

Number of correct predictions

Total number of predictions

#### Let's Look at the Code!

How do we compute the accuracy of solving analogies in a test set?