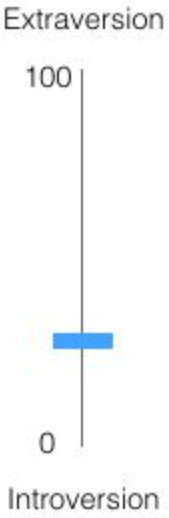
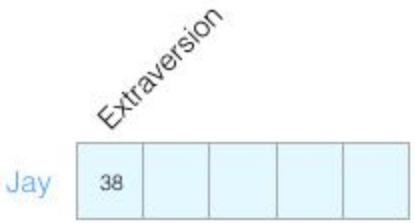
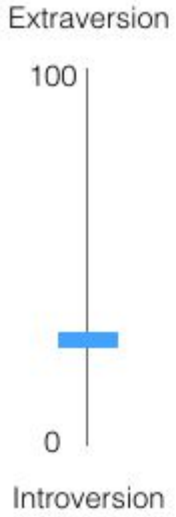
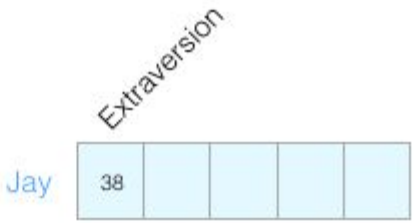
Introducere în NLP

Zăvelcă Miruna-Andreea
miruna-andreea.zavelca@unibuc.ro
Universitatea din București

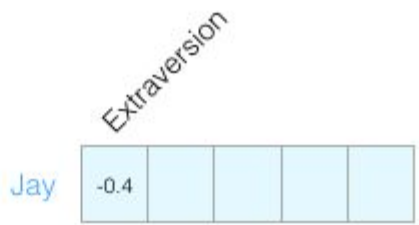

Extraversion
100
0
Introversion

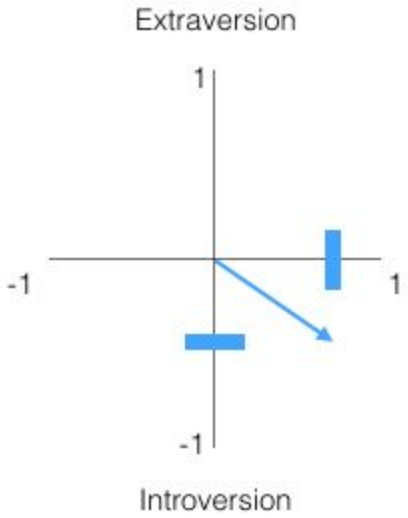

Extraversion
Jay
38


Extraversion
100
0
Introversion


Extraversion
Jay
38

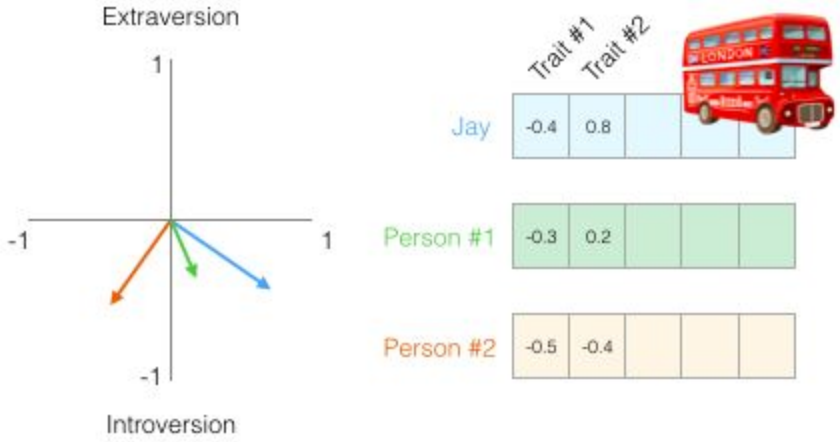

Extraversion
1
-1
Introversion


Extraversion
Jay
-0.4

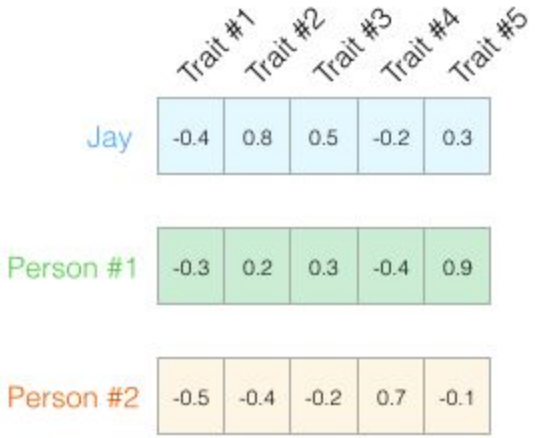


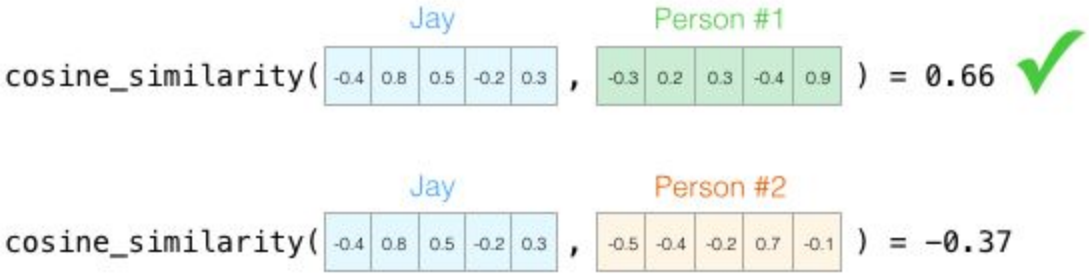
	Trait #1	Trait #2			
Jay	-0.4	0.8			

Openness to experience	79 out of 100
Agreeableness	75 out of 100
Conscientiousness	42 out of 100
Negative emotionality	50 out of 100
Extraversion	58 out of 100


Extraversion
1
-1
1
-1
Introversion
Trait #1
Trait #2
Jay
-0.4
0.8
Person #1
-0.3
0.2
Person #2
-0.5
-0.4


Jay
Person #1
cosine_similarity(-0.4 0.8 , -0.3 0.2) = 0.87
Jay
Person #2
cosine_similarity(-0.4 0.8 , -0.5 -0.4) = -0.20


Trait #1
Trait #2
Trait #3
Trait #4
Trait #5
Jay
-0.4
0.8
0.5
-0.2
0.3
Person #1
-0.3
0.2
0.3
-0.4
0.9
Person #2
-0.5
-0.4
-0.2
0.7
-0.1


Jay
Person #1
cosine_similarity(-0.4 0.8 0.5 -0.2 0.3 , -0.3 0.2 0.3 -0.4 0.9) = 0.66
Jay
Person #2
cosine_similarity(-0.4 0.8 0.5 -0.2 0.3 , -0.5 -0.4 -0.2 0.7 -0.1) = -0.37

1- We can represent things (and people) as vectors of numbers
(Which is great for machines!)

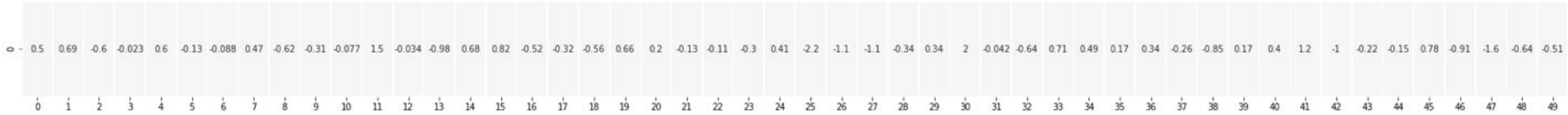
Jay	-0.4	0.8	0.5	-0.2	0.3

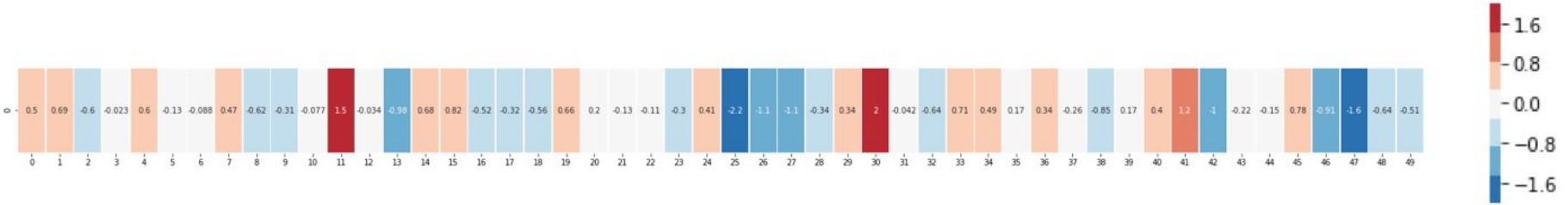
2- We can easily calculate how similar vectors are to each other

The people most similar to Jay are:

	cosine_similarity ▼
Person #1	0.86
Person #2	0.5
Person #3	-0.20

Reprezentarea cuvintelor

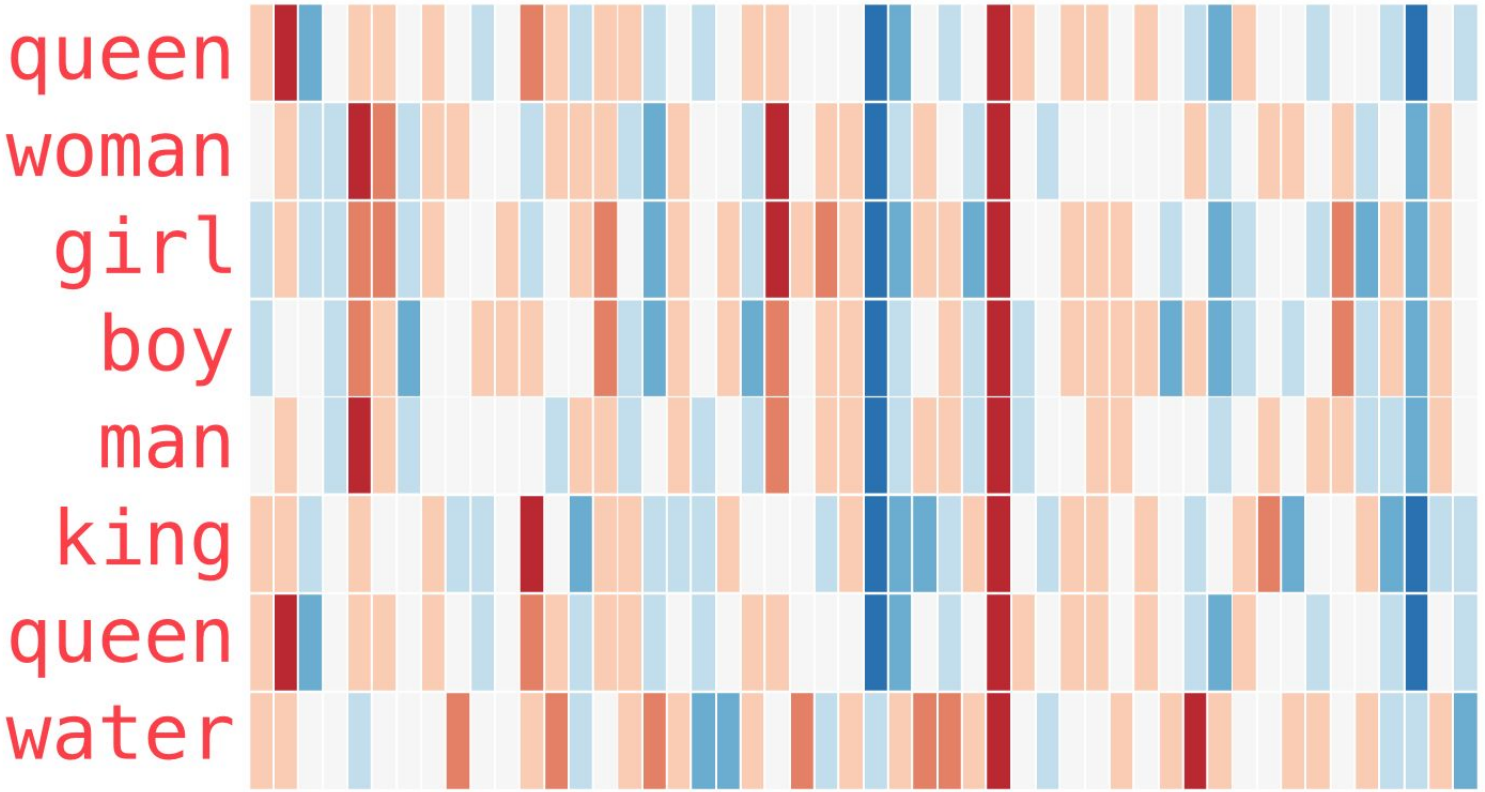

0
0.5 0.69 -0.6 -0.023 0.6 -0.13 -0.088 0.47 -0.62 -0.31 -0.077 1.5 -0.034 -0.98 0.68 0.82 -0.52 -0.32 -0.56 0.66 0.2 -0.13 -0.11 -0.3 0.41 -2.2 -1.1 -1.1 -0.34 0.34 2 -0.042 -0.64 0.71 0.49 0.17 0.34 -0.26 -0.85 0.17 0.4 1.2 -1 -0.22 -0.15 0.78 -0.91 -1.6 -0.64 -0.51
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44 45 46 47 48 49


0
0.5 0.69 -0.6 -0.023 0.6 -0.13 -0.088 0.47 -0.62 -0.31 -0.077 1.5 -0.034 -0.98 0.68 0.82 -0.52 -0.32 -0.56 0.66 0.2 -0.13 -0.11 -0.3 0.41 -2.2 -1.1 -1.1 -0.34 0.34 2 -0.042 -0.64 0.71 0.49 0.17 0.34 -0.26 -0.85 0.17 0.4 1.2 -1 -0.22 -0.15 0.78 -0.91 -1.6 -0.64 -0.51
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44 45 46 47 48 49
1.6
0.8
0.0
−0.8
−1.6

“king”

“Man”

“Woman”


queen
woman
girl
boy
man
king
queen
water

king - man + woman ~= queen

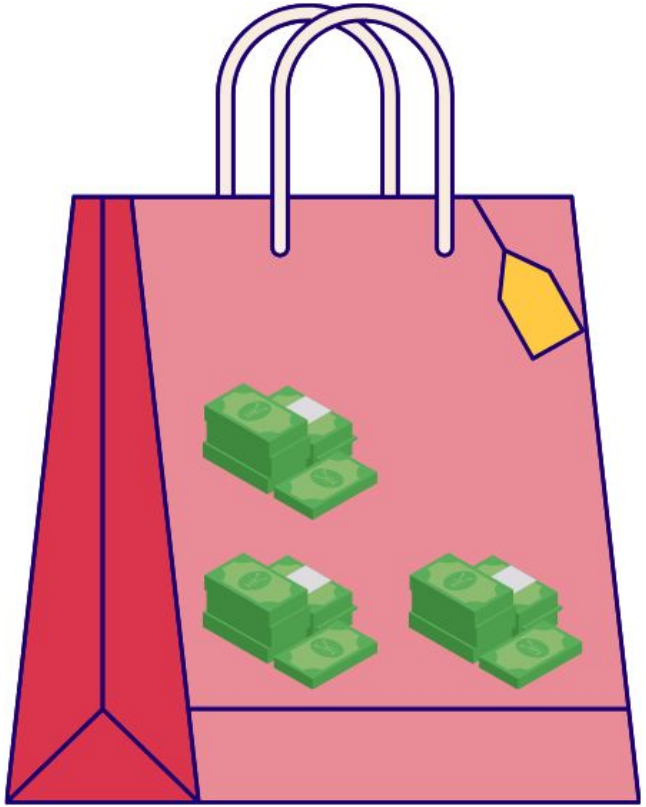
```
model.most_similar(positive=["king","woman"], negative=["man"])
```

```
[('queen', 0.8523603677749634),
 ('throne', 0.7664333581924438),
 ('prince', 0.7592144012451172),
 ('daughter', 0.7473883032798767),
 ('elizabeth', 0.7460219860076904),
 ('princess', 0.7424570322036743),
 ('kingdom', 0.7337411642074585),
 ('monarch', 0.721449077129364),
 ('eldest', 0.7184862494468689),
 ('widow', 0.7099430561065674)]
```

Metode de Reprezentare

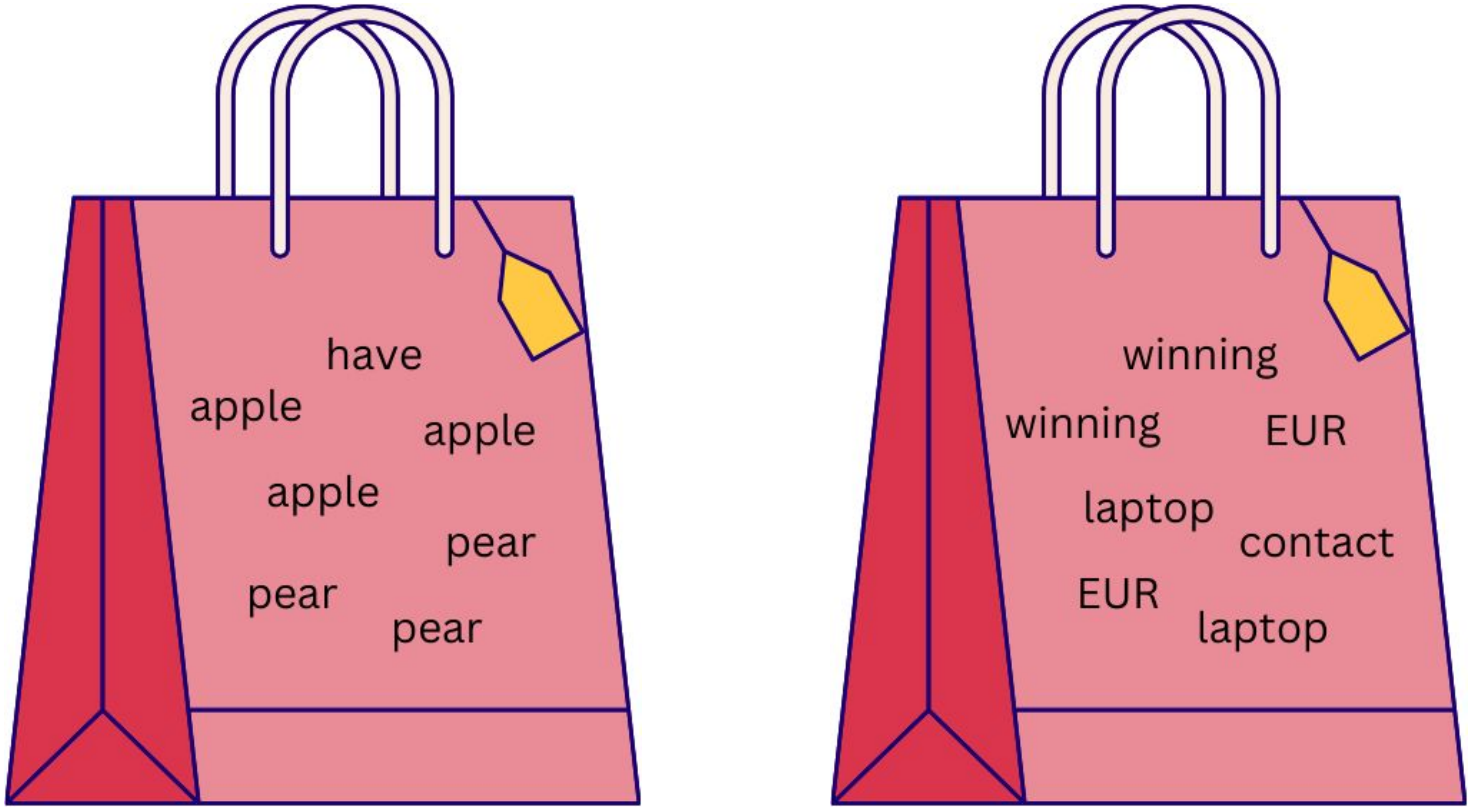
Bag of ?

Bag of ?



Bag of ?

Bag of Words


have
apple
apple
apple
pear
pear
pear
winning
winning
EUR
laptop
contact
EUR
laptop

Bag of Words pentru Spam Detection

Subject: Prize Notification

Microsoft Iberica S.L Lottery Intl. Program FOREIGN SERVICE SECTION BARCELONA. REFERENCE NUMBER:YUKFQ/RYYHJ
BATCH NUMBER: 2016/WTN
OFFICIAL WINNING NOTIFICATION.
We are pleased to inform you of the released results of the Microsoft Iberica S.L Sweepstakes Promotion in conjunction with foundations for the promotion of software products organized for Software users. This Program was held in Barcelona-Spain; Wherein your email address emerged as one of the online Winning emails in the 1st category and therefore attracted a cash award of EUR344,000.00 and a Mac laptop/iPhone. Your laptop, certificate of winnings and your cheque of (EUR344,000.00) will be sent to your contact address in your location. To file for claims of the release of your winnings, contact the Customer Service Officer with the information below:

FULL NAMES, ADDRESS, SEX, AGE, MARITAL STATUS, OCCUPATION, TELEPHONE NUMBER, COUNTRY, BATCH NUMBER, REFERENCE NUMBER: Email: cuservdept@excite.co.jp Contact Person: Manuel Vizner [CSO]

Also, please, fill out our customer satisfaction survey at www.excite.co.jp/Survey.aspx?s=4674c60&surv_id=DNTY

Congratulations!!

Sincerely,
Mrs. Miriam Inaki
Online Coordinator

Bag of Words pentru Spam Detection

Subject: Prize Notification

Microsoft Iberica S.L Lottery Intl. Program FOREIGN SERVICE SECTION BARCELONA. REFERENCE NUMBER:YUKFQ/RYYHJ
BATCH NUMBER: 2016/WTN
OFFICIAL WINNING NOTIFICATION.
We are pleased to inform you of the released results of the Microsoft Iberica S.L Sweepstakes Promotion in conjunction with foundations for the promotion of software products organized for Software users. This Program was held in Barcelona-Spain; Wherein your email address emerged as one of the online Winning emails in the 1st category and therefore attracted a cash award of EUR344,000.00 and a Mac laptop/iPhone. Your laptop, certificate of winnings and your cheque of (EUR344,000.00) will be sent to your contact address in your location. To file for claims of the release of your winnings, contact the Customer Service Officer with the information below:

FULL NAMES, ADDRESS, SEX, AGE, MARITAL STATUS, OCCUPATION, TELEPHONE NUMBER, COUNTRY, BATCH NUMBER, REFERENCE NUMBER: Email: cuservdept@excite.co.jp Contact Person: Manuel Vizner [CSO]

Also, please, fill out our customer satisfaction survey at www.excite.co.jp/Survey.aspx?s=4674c60&surv_id=DNTY

Congratulations!!

Sincerely,
Mrs. Miriam Inaki
Online Coordinator

Bag of Words pentru Spam Detection

Subject: Prize Notification

Microsoft Iberica S.L Lottery Intl. Program FOREIGN SERVICE SECTION BARCELONA. REFERENCE NUMBER:YUKFQ/RYYHJ
BATCH NUMBER: 2016/WTN
OFFICIAL WINNING NOTIFICATION.
We are pleased to inform you of the released results of the Microsoft Iberica S.L Sweepstakes Promotion in conjunction with foundations for the promotion of software products organized for Software users. This Program was held in Barcelona-Spain; Wherein your email address emerged as one of the online Winning emails in the 1st category and therefore attracted a cash award of EUR344,000.00 and a Mac laptop/iPhone. Your laptop, certificate of winnings and your cheque of (EUR344,000.00) will be sent to your contact address in your location. To file for claims of the release of your winnings, contact the Customer Service Officer with the information below:

FULL NAMES, ADDRESS, SEX, AGE, MARITAL STATUS, OCCUPATION, TELEPHONE NUMBER, COUNTRY, BATCH NUMBER, REFERENCE NUMBER: Email: cuservdept@excite.co.jp Contact Person: Manuel Vizner [CSO]

Also, please, fill out our customer satisfaction survey at www.excite.co.jp/Survey.aspx?s=4674c60&surv_id=DNTY

Congratulations!!

Sincerely,
Mrs. Miriam Inaki
Online Coordinator

winning	3
EUR	2
laptop	2
Barcelona	2
email	2
contact	2
...	...

Bag of Words în cod

	winning	EUR	laptop	Barcelona	email	contact	...
D1	3	2	2	2	2	2	...
D2	2	1	1	0	0	1	...
D3	1	0	0	0	0	1	...

Term Frequency - Inverse Document Frequency

Ce căutăm?

1

2

3

4

TFIDF - Term Frequency Inverse Document Frequency

TFIDF = TF * IDF

$$\text{TF (cuvânt, document)} = \frac{\text{număr apariții cuvânt în document}}{\text{număr de cuvinte în document}}$$

$$\text{IDF (cuvânt, documente)} = \frac{\text{număr de documente în corpus}}{\text{număr de documente care conțin cuvântul dat}}$$

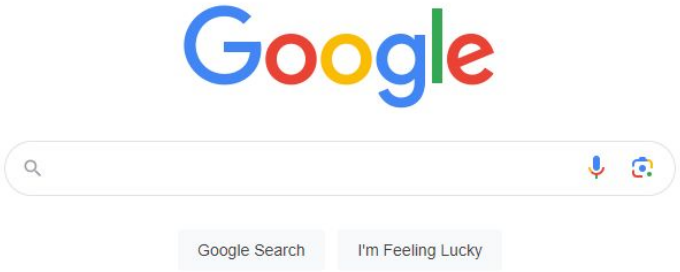
TFIDF - Calculam ce relevanta are un obiect intr-un document raportat la tot setul de documente

	1.1	1.3	1.1	0
	0.7	0	0	0
	0	0	0.7	2

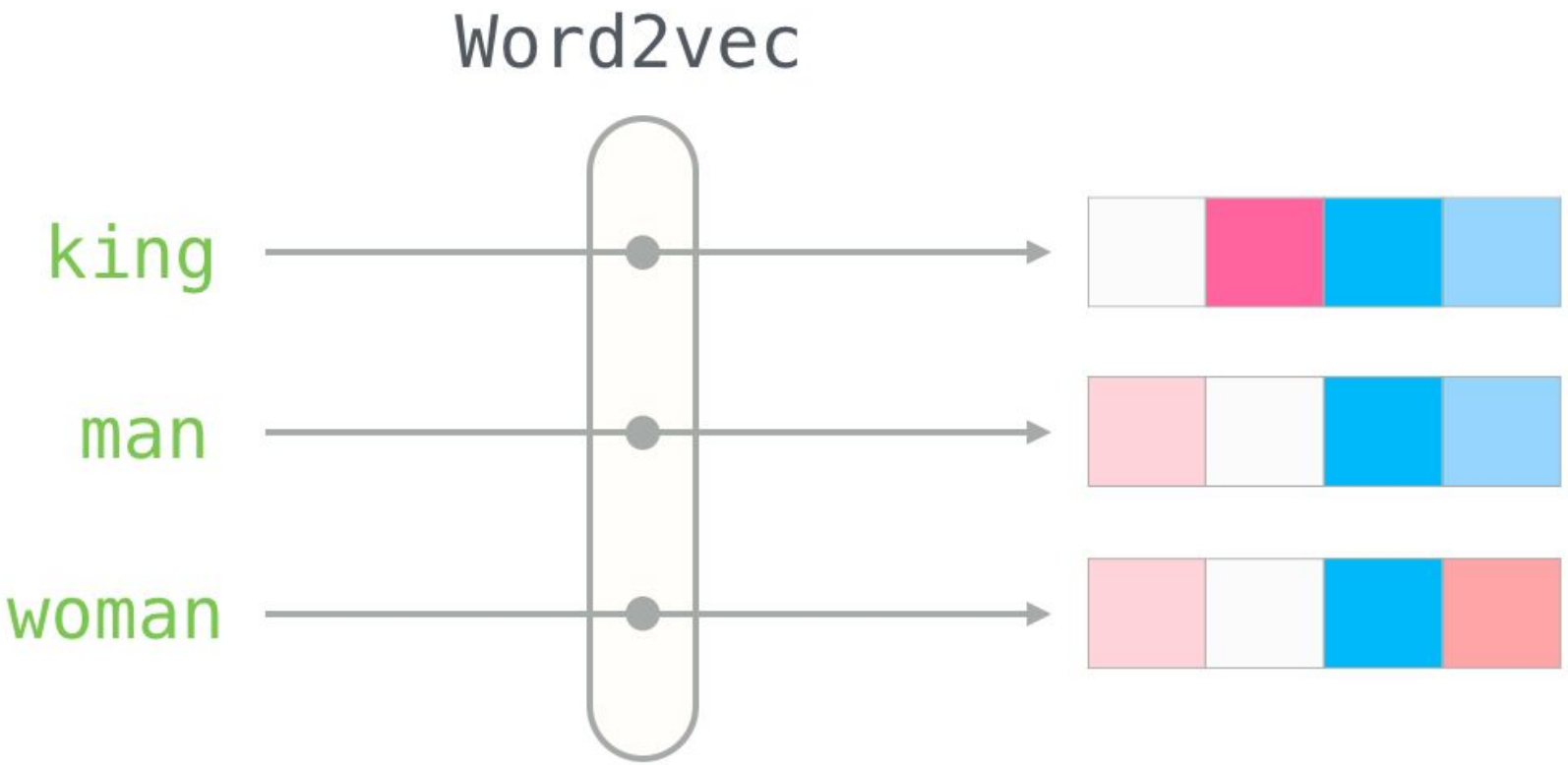
TFIDF în cod

	fructe	arme	bani
D1	1,1	0,7	0
D2	1,3	0	0
D3	1,1	0	0,7
D4	0	0	2

TFIDF pentru search engines



Word2Vec


Word2vec
king
man
woman


Thou shalt
SEND
1
+
is
not
I
1 2 3 4 5 6 7 8 9 0
q w e r t y u i o p
@ # & * - + = ( )
a s d f g h j k l
_ $ " ' : ; / !
z x c v b n m
123
English (US)
,!?
.

Thou shalt not make a machine in the likeness of a human mind

Sliding window across running text

thou	shalt	not	make	a	machine	in	the	…

Dataset

input 1	input 2	output

Thou shalt not make a machine in the likeness of a human mind

Sliding window across running text

thou	shalt	not	make	a	machine	in	the	…

Dataset

input 1	input 2	output
thou	shalt	not

Thou shalt not make a machine in the likeness of a human mind

Sliding window across running text

thou	shalt	not	make	a	machine	in	the	…
thou	shalt	not	make	a	machine	in	the	

Dataset

input 1	input 2	output
thou	shalt	not
shalt	not	make

Thou shalt not make a machine in the likeness of a human mind

Sliding window across running text

thou	shalt	not	make	a	machine	in	the	…
thou	shalt	not	make	a	machine	in	the	
thou	shalt	not	make	a	machine	in	the	
thou	shalt	not	make	a	machine	in	the	
thou	shalt	not	make	a	machine	in	the	

Dataset

input 1	input 2	output
thou	shalt	not
shalt	not	make
not	make	a
make	a	machine
a	machine	in

Jay was hit by a _____

Jay was hit by a _____ bus

Jay was hit by a _____ bus in...

by	a	red	bus	in

Jay was hit by a _____ bus in...

input 1	input 2	input 3	input 4	output
by	a	bus	in	red

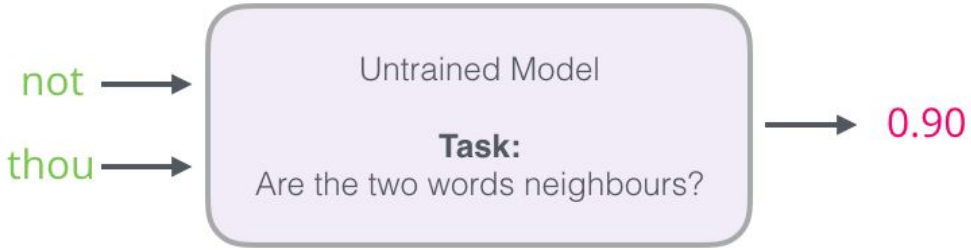
Jay was hit by a red bus in...

by	a	red	bus	in

input	output
red	by
red	a
red	bus
red	in

Change Task from

To:



Q&A

Images from Jay Alammar