



Word-Embedding et sentiments des ménages avec Twitter

KIM ANTUNEZ, ROMAIN LESAUVAGE ET ALAIN
QUARTIER-LA-TENTE
11/06/2020
Ensaë — 2019-2020

Sommaire

1. Introduction

2. *word2vec*

3. Évaluation du modèle

Introduction

Sommaire

1. Introduction

2. *word2vec*

3. Évaluation du modèle

Sommaire

1. Introduction

2. *word2vec*

3. Évaluation du modèle

Évaluation sur le corpus de tweets (1/2)

« Notre » modèle

Spearman : 0,57 (p-v : 4,1 %)

➡ **bons résultats**

bonjour (669)	femme (264)	1 (765)	samedi (203)
😊 (0,59)	quelle (0,49)	5 (0,55)	soir (0,57)
😊 (0,59)	cette (0,46)	mois (0,51)	vivement (0,51)
merci (0,54)	une (0,44)	10 (0,49)	demain (0,50)
nuit (0,48)	vie (0,44)	2 (0,48)	end (0,48)
bisous (0,47)	grippe (0,44)	top (0,48)	weekend (0,47)
bonne (0,47)	belle (0,43)	depuis (0,47)	matin (0,45)
😞 (0,46)	ma (0,43)	saison (0,46)	jeudi (0,45)
vous (0,46)	magnifique (0,43)	ans (0,44)	prochain (0,43)
plaisir (0,44)	nouvelle (0,43)	jours (0,43)	week (0,43)
allez (0,43)	vidéo (0,39)	3 (0,43)	🌸 (0,42)

$ep = 80 / w = 4 / lr = 0,02 / dim = 100 / base : 100\ 000\ tweets$

Évaluation sur le corpus de tweets (1/2)

« Notre » modèle

Spearman : 0,57 (p-v : 4,1 %)

➡ **bons résultats**

bonjour (669)	femme (264)	1 (765)	samedi (203)
👋 (0,59)	quelle (0,49)	5 (0,55)	soir (0,57)
😊 (0,59)	cette (0,46)	mois (0,51)	vivement (0,51)
merci (0,54)	une (0,44)	10 (0,49)	demain (0,50)
nuit (0,48)	vie (0,44)	2 (0,48)	end (0,48)
bisous (0,47)	grippe (0,44)	top (0,48)	weekend (0,47)
bonne (0,47)	belle (0,43)	depuis (0,47)	matin (0,45)
😞 (0,46)	ma (0,43)	saison (0,46)	jeudi (0,45)
vous (0,46)	magnifique (0,43)	ans (0,44)	prochain (0,43)
plaisir (0,44)	nouvelle (0,43)	jours (0,43)	week (0,43)
allez (0,43)	vidéo (0,39)	3 (0,43)	🌸 (0,42)

$ep = 80 / w = 4 / lr = 0,02 / dim = 100 / base : 100\ 000\ tweets$

Modèle Gensim

Spearman : 0,50 (p-v : 0,0 %)

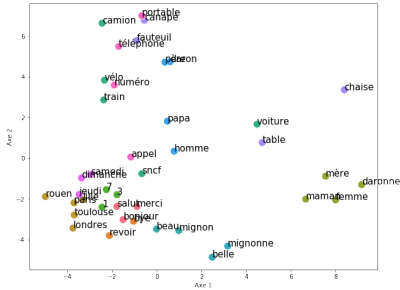
➡ **très bons résultats**

bonjour (17 043)	femme (6 177)	1 (21 055)	samedi (4 917)
bonsoir (0,85)	filles (0,86)	2 (0,65)	vendredi (0,88)
bjr (0,75)	copine (0,74)	3 (0,64)	jeudi (0,86)
hello (0,71)	meuf (0,71)	6 (0,63)	lundi (0,83)
salut (0,66)	demoiselle (0,66)	4 (0,62)	mercredi (0,83)
coucou (0,55)	nana (0,66)	7 (0,60)	dimanche (0,83)
transmets (0,49)	nièce (0,66)	5 (0,58)	mardi (0,76)
désagrement (0,48)	sœur (0,65)	9 (0,58)	demain (0,72)
avezvous (0,48)	barbe (0,65)	8 (0,56)	barathon (0,56)
bettembourg (0,48)	maman (0,64)	1e (0,55)	22h45 (0,55)
hey (0,47)	princesse (0,64)	34 (0,53)	20h (0,54)

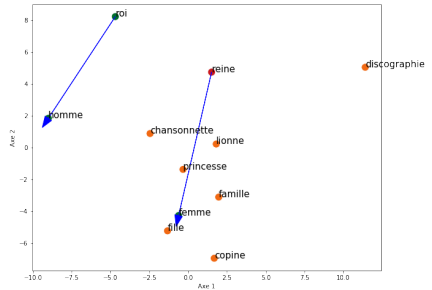
$ep = 100 / w = 4 / lr = 0,02 / dim = 100 / base : ensemble\ des\ tweets$

10 plus proches voisins par similarité cosinus

Évaluation sur le corpus de tweets (2/2)



ACP sur un corpus réduit de mots.



$$\overrightarrow{Roi} - \overrightarrow{Homme} + \overrightarrow{Femme} = ?$$

➡ Réduction de dimension des vecteurs-mots et (parfois) opérations sur les mots **convaincants**

Merci pour votre attention

 ARKEnsaë/TweetEmbedding

 Rapport du projet

