Universite Paris-Saclay

Text-Mining

Word Embedding

Word2Vec Word Embedding training using Gensim and Fasttext.

Pawel Budzynski

September 10, 2022

1 Introduction

In this exercise two implementations of word embedding model are going to be used together with two corpuses including text in French language. The goal is to (1) produce word embedding that makes sense from the language perspective (2) inspect differences between techniques and (3) investigate the impact of the data used for training on the outcome. The used models are Gensim Word2Vec skipgram, Gensim Word2Vec CBOW and Fasttext CBOW. Two corpuses are used: small one with medical text FrenchMed Corpus and big one with non-medical text FrenchPress Corpus (https://quaerofrenchmed.limsi.fr/).

2 Models trained on FrenchMed Corpus

In most of the cases the closes words are not of similar meanings but seem more like words that occurs close to target words in a text. For example, for a word "patient" most of the close words refer to a person state and treatment application. It is visible that models quite successfully captured a meaning of words *traitement*, *maladie* and *solution* as their closest words represents types of treatments, diseases and solutions respectively. Closest words for *jaune* seem to not have much sense, the best shot is "orange". That is for sure caused by the data type used for training. Strictly medical terms are grouped properly while general words like colors did not have enough of representatives hence their similarity to other words was not well recognised.

Results of Gensim Skipgram and CBOW are somehow similar with the order of similar words shuffled. The results of Fasttext differ significantly however it might be because of the differences in pre-processing of the text. It is clear that Fasttext implementation applied less of text transformation hence typos/plural forms/forms of a word are present, for example *patient*, *patiente*, *patients*, *Patient*. In the last part of the report I will apply Fasttext to a text pre processed with Gensim utils. In the results above I would say that results produced by Gensim models looks more meaningful to me. On the other hand Fasttext was able to capture obvious similarities successfully which may give a clue about correctness of the training.

Table 1: Closest neighbors for word *patient*.

Gensim Skipgram	Gensim CBOW	Fasttext CBOW
stimulateur	Montrez	Patient
repos	carte	parvient
encourus	alerte	maintient
souffre	attentif	recevaient
gériatriques	souffre	avaient
rencontrés	soigneusement	aient
certitude	évocateurs	soient
trouverez	certitude	Contient
Carte	symptômes	gradient
déterminer	afin	excipient

Table 2: Closest neighbors for word *traitement*.

Gensim Skipgram	Gensim CBOW	Fasttext CBOW
Reprise	être	Traitement
le	doit	traitment
Thoraciques	patients	Taaitement
agrafage	par	Allaitement
définitif	Tasmar	étroitement
volets	médecin	allaitement
aphtes	instauré	évitement
péricardite	prise	immédiatement
paralysies	détecter	recrutement
spécialistes	nécessité	correctement

Table 3: Closest neighbors for word *maladie*.

Gensim Skipgram	Gensim CBOW	Fasttext CBOW
Parkinson	Parkinson	Maladie
AINS	liée	malade
Légionnaires	Crohn	maldi
mouton	avancé	malgré
Inflammation	Légionnaires	malin
Polyneuropathie	stabiliser	maïs
constituée	SIDA	Parkinson
stabiliser	atteint	avancée
vraie	Hirsprung	avancé
Marfan	Bourneville	maladies

Table 4: Closest neighbors for word *solution*.

Gensim Skipgram	Gensim CBOW	Fasttext CBOW
Ajoutez	ml	Dissolution
Lepirudine	injectable	Solution
Voie	diluer	dilution
aseptique	perfusion	Pollution
préparée	mg	evolution
rincez	Chaque	Evolution
Poudre	contient	déglutition
Toute	dosée	Substitution
préparer	poudre	évolution
dosée	Débit	exécution

Table 5: Closest neighbors for word *jaune*.

Gensim Skipgram	Gensim CBOW	Fasttext CBOW
pâle	pâle	hexane
Calotermes	Calotermes	routine
Fabr	Fabr	bobine
flavicollis	flavicollis	Triacétine
orange	orange	titane
hexagonaux	oxyde	triacétine
anormale	mosaïque	gêne
talc	talc	machine
termite	Méthylhydroxypropylcellulose	Lane
navet	Talc	aptine

3 Models trained on FrenchPress Corpus

On the contrary to the medical corpus the results that come from non-medical corpus seems to be correct however less specific. That is of course caused by the domain of the language used. The Press corpus is also bigger one thus it is not sure whether the amount of training applied was enough to train the models properly. The text base is clearly less medical hence not all similar words have strictly medical character like in case of *traitement* or *solution* some of close words come from non-medical domain but still make sense.

Table 6: Closest neighbors for word *patient*.

Gensim Skipgram	Gensim CBOW	Fasttext CBOW
infantilisé	hospitalisé	patientent
hospitalisé	contaminé	impatient
cancéreux	cancéreux	patiente
soignant	livré	patients
extraire	ricane	impatientent
humble	infantilisé	patiemment
palliatifs	soignant	impatiente
polluée	séropositif	patientera
bas-âge	déboutés	patienter
interpénétrer	interpénétrer	patience

Table 7: Closest neighbors for word *traitement*.

Gensim Skipgram	Gensim CBOW	Fasttext CBOW
gériatrie	gériatrie	traitements
anti-douleur	antidouleur	retraitement
Ahivor	inégalité	maladroitement
asservissement	cohérent	subitement
Palmade	générateurs	étroitement
Lariboisière	médicamenteux	bêtement
Bergman	anti-douleur	traite
impose-t-elle	remédier	traiter
médicamenteux	royalties	dépècement
asilah	Stöhr	traitent

Table 8: Closest neighbors for word *maladie*.

Gensim Skipgram	Gensim CBOW	Fasttext CBOW
neurologique	Alzheimer	maladies
Parkinson	transmissible	malade
Alzheimer	dingue	maladroit
pulmonaire	épidémie	malades
succombent	cancéreux	malawite
transmissible	disséminer	rhinovirus
orpheline	maladies	Mladic
souffrait	virale	maladresse
virale	gènes	maladroite
161	pneumonie	maladroitement

Table 9: Closest neighbors for word *solution*.

Gensim Skipgram	Gensim CBOW	Fasttext CBOW
pacifique	pacifique	résolution
cochonneries	sodium	dissolution
lancinant	mesure	dilution
Hyperion	stock	solutions
garantissant	cochonneries	évolution
carbure	acceptable	résolutions
recadrage	consensuelle	caution
sodium	parvenir	pollution
constructifs	lancinant	révolution
agréée	expropriation	vexation

Table 10: Closest neighbors for word *jaune*.

Gensim Skipgram	Gensim CBOW	Fasttext CBOW
maillot	Saâdoune	jaune-vert
emparé	maillot	aune
Pena	emparé	Neptune
Nazon	grenadine	Jeune
grenadine	Pena	dune
Saâdoune	Abdellatif	jaunes
Bradeley	décaleront	faune
endossant	Bradley	Saâdoune
Jandoubi	Âge	l'une
Baden	Nazon	Lachhab

4 Impact of the data

The impact of the data used was discussed partially in the previous parts. Direct comparisons of the same model outcomes given different data appears to be hard for not a fluent French speaker. What is clear that similar words from medical corpus seems to be coming from medical domain while for press corpus the terms are more general. That might be caused by both, wider range of vocabulary used in the press corpus as well as lack of many specific medical terms in this corpus.

4.1 Comparison for Word2Vec Skipgram

Table 11: Closest neighbors for word *patient*.

FrenchMed	FrenchPress
stimulateur	infantilisé
repos	hospitalisé
encourus	cancéreux
souffre	soignant
gériatriques	extraire
rencontrés	humble
certitude	palliatifs
trouverez	polluée
Carte	bas-âge
déterminer	interpénétrer

Table 12: Closest neighbors for word *traitement*.

FrenchMed	FrenchPress
Reprise	gériatrie
le	anti-douleur
Thoraciques	Ahivor
agrafage	asservissement
définitif	Palmade
volets	Lariboisière
aphtes	Bergman
péricardite	impose-t-elle
paralysies	médicamenteux
spécialistes	asilah

Table 13: Closest neighbors for word *maladie*.

FrenchMed	FrenchPress
Parkinson	neurologique
AINS	Parkinson
Légionnaires	Alzheimer
mouton	pulmonaire
Inflammation	succombent
Polyneuropathie	transmissible
constituée	orpheline
stabiliser	souffrait
vraie	virale
Marfan	161

Table 14: Closest neighbors for word *solution*.

FrenchMed	FrenchPress
Ajoutez	pacifique
Lepirudine	cochonneries
Voie	lancinant
aseptique	Hyperion
préparée	garantissant
rincez	carbure
Poudre	recadrage
Toute	sodium
préparer	constructifs
dosée	agréée

Table 15: Closest neighbors for word *jaune*.

FrenchMed	FrenchPress
pâle	maillot
Calotermes	emparé
Fabr	Pena
flavicollis	Nazon
orange	grenadine
hexagonaux	Saâdoune
anormale	Bradeley
talc	endossant
termite	Jandoubi
navet	Baden

4.2 Comparison for Word2Vec CBOW

Closest neighbors for: patient

FrenchMed	FrenchPress
Montrez	hospitalisé
carte	contaminé
alerte	cancéreux
attentif	livré
souffre	ricane
soigneusement	infantilisé
évocateurs	soignant
certitude	séropositif
symptômes	déboutés
afin	interpénétrer

Closest neighbors for: traitement

FrenchMed	FrenchPress
être	gériatrie
doit	antidouleur
patients	inégalité
par	cohérent
Tasmar	générateurs
médecin	médicamenteux
instauré	anti-douleur
prise	remédier
détecter	royalties
nécessité	Stöhr

Closest neighbors for: maladie

FrenchMed	FrenchPress
Parkinson	Alzheimer
liée	transmissible
Crohn	dingue
avancé	épidémie
Légionnaires	cancéreux
stabiliser	disséminer
SIDA	maladies
atteint	virale
Hirsprung	gènes
Bourneville	pneumonie

Closest neighbors for: solution

FrenchMed	FrenchPress
ml	pacifique
injectable	sodium
diluer	mesure
perfusion	stock
mg	cochonneries
Chaque	acceptable
contient	consensuelle
dosée	parvenir
poudre	lancinant
Débit	expropriation

Table 16: Closest neighbors for word *jaune*.

FrenchMed	FrenchPress
pâle	Saâdoune
Calotermes	maillot
Fabr	emparé
flavicollis	grenadine
orange	Pena
oxyde	Abdellatif
mosaïque	décaleront
talc	Bradley
Méthylhydroxypropylcellulose	Âge
Talc	Nazon

4.3 Comparison for Fasttext CBOW

Table 17: Closest neighbors for word *patient*.

FrenchMed	FrenchPress
Patient	patientent
parvient	impatient
maintient	patiente
recevaient	patients
avaient	impatientent
aient	patiemment
soient	impatiente
Contient	patientera
gradient	patienter
excipient	patience

Table 18: Closest neighbors for word *traitement*.

FrenchMed	FrenchPress
Traitement	traitements
traitment	retraitement
Taaitement	maladroitement
Allaitement	subitement
étroitement	étroitement
allaitement	bêtement
évitement	traite
immédiatement	traiter
recrutement	dépècement
correctement	traitent

Table 19: Closest neighbors for word *maladie*.

FrenchMed	FrenchPress
Maladie	maladies
malade	malade
maldi	maladroit
malgré	malades
malin	malawite
maïs	rhinovirus
Parkinson	Mladic
avancée	maladresse
avancé	maladroite
maladies	maladroitement

Table 20: Closest neighbors for word *solution*.

FrenchMed	FrenchPress
Dissolution	résolution
Solution	dissolution
dilution	dilution
Pollution	solutions
evolution	évolution
Evolution	résolutions
déglutition	caution
Substitution	pollution
évolution	révolution
exécution	vexation

Table 21: Closest neighbors for word *jaune*.

FrenchMed	FrenchPress
hexane	jaune-vert
routine	aune
bobine	Neptune
Triacétine	Jeune
titane	dune
triacétine	jaunes
gêne	faune
machine	Saâdoune
Lane	l'une
aptine	Lachhab

5 Models trained with pre-processed FrenchMed corpus

In this part I tried to fix the problem of Fasttext results where the similar words are in fact the same words transformed or mistyped. For that, I have applied "genism.utils.simple_preprocess()" to the medical corpus and saved the pre processed text to a new file. It is still unclear whether gensim applied some additional data cleaning or their models fail to find obvious connections. For example, *patient* and *patients* are close in Fasttext model but not in the others. After all, after applying the pre-processing to the data the results of Fasttext look to be more generalsed and less repetitions are present.

Table 22: Closest neighbors for word *patient*.

Gensim Skipgram	Gensim CBOW	Fasttext CBOW
stimulateur	montrez	parvient
encourus	carte	avaient
attentif	alerte	ajoutent
souffre	attentif	maintient
remise	remise	aient
courants	souffre	recevaient
repos	devra	soient
certitude	certitude	doivent
existante	éliminer	souvent
montrez	que	gradient

Table 23: Closest neighbors for word *traitement*.

Gensim Skipgram	Gensim CBOW	Fasttext CBOW
le	instauré	traitment
par	surveillé	taaitement
du	approprié	allaitement
dans	habitué	étroitement
patients	minimum	évitement
de	expérimenté	immédiatement
tacrolimus	arrêté	avortement
fk	doit	fortement
semaines	tasmar	lentement
et	réévaluer	directement

Table 24: Closest neighbors for word *jaune*.

Gensim Skipgram	Gensim CBOW	Fasttext CBOW
orange	pâle	hexane
pâle	orange	chaîne
fabr	calotermes	machine
flavicollis	flavicollis	chine
hexagonaux	fabr	bobine
calotermes	navet	routine
replicase	mosaïque	titane
mosaïque	talc	triacétine
anormale	hexagonaux	celgene
navet	anormale	levane

Table 25: Closest neighbors for word *solution*.

Gensim Skipgram	Gensim CBOW	Fasttext CBOW
ajoutez	ajoutez	dissolution
lepirudine	diluer	dilution
reconstituée	dosée	pollution
préparée	préparée	evolution
transparente	ml	déglutition
diluée	lepirudine	distribution
perfusable	injectable	évolution
poudre	reconstituée	substitution
dosée	contient	microdélétion
sucre	dissoudre	redistribution

Table 26: Closest neighbors for word *maladie*.

Gensim Skipgram	Gensim CBOW	Fasttext CBOW
kahler	parkinson	malade
waldenstroem	crohn	malt
iconographique	polyneuropathie	maldi
parkinson	légionnaires	malgré
wolman	kahler	malin
lobstein	basedow	mao
légionnaires	bourneville	malta
polyneuropathie	médiocre	avancé
ains	habitué	maïs
hirsprung	recklinghausen	malherbe

6 Conclusions

In this exercise 3 kinds of word embedding models were trained using 2 kinds of data. However, the implementations of the models differ, it is also hard to judge about their quality since even running the training of the same model with the same data may produce different results eventually. In my opinion, getting satisfying results required more hyperparameters tuning in case of gensim implementations. The corpus used for training seems to have more direct impact on the results and a lot depends on the text and its pre-processing. Before training and using word embedding a person should ask themselves what kind of text they are going to work with and what are the specific needs for the particular application. For example, to build a model that operates on medical text the corpus containing sufficient amount of medical terms should be used.