Application :

Analyse de sentiment

Traduction

Reconnaissance vocale

Pourquoi adapter au français ?

jamais fait actuellement Algo anglais appliqué

Pas adapté ex : l', é', 2ème et 3ème personne

~~Nuit du~~ Introduction } Nos sommes ici

* Ex plus simple d'utilisation.

~~Des~~, la compréhension de notre langage (c-a-d le français) par les ordinateurs est de plus en plus important ; car ils sont de plus en plus présent dans notre vie ~~(Ex Smartphone)~~ et le seront de + en +, ex : Smartphone, Voiture Autonome, Siri".

Quelques.

Le français dans l'ère du "Deep Learning" et les "Word Embeddings"

Ce projet consiste en l'utilisation de techniques connes avec le nom de "Word Embeddings" pour le français. Ces techniques ont été largement étudiées pour l'anglais [1][2][3][4] et récemment pour des autres langues. Normalement, les auteurs de cette technique ont démontré qu'elle avait une meilleur performance par des évaluations des analogies [3][4]. C'est-à-dire, qu'une méthode était considérée comme supérieure si elle arrivée à trouver correctement un des mots d'une analogie avec 4 mots. Par exemple, la tâche consiste à trouver correctement *Queen* sachant qu'on connaît *Man*, *King* et *Woman*. Presque 20 mil exemples sont données par les auteur de cette technique [3]. Cependant, l'évaluation est appliqué que pour l'anglais et complètement inexistant pour le français ou autres langues. Également, la source avec laquelle la méthode est entraîné est sans aucune doute un ressource précieux pour trouver de bons résultats. Pour l'anglais, la Wikipédia a été utilisé avec de bons résultats. Pour le français, on se content d'utiliser la collection objective et pas une collection générique afin de comprendre les caractéristiques particulières de la langue. Probablement la manque d'une ressource d'évaluation est à l'origine de cette situation.

Les principaux objectives de ce projet sont :

- La traduction et analyses des collection d'évaluation par analogie existants pour l'évaluation de words embeddings pour la langue française.
- La construction de modèles avec les méthodes existants et de ressources ouverts pour la langue française.
- L'évaluation et comparaison de différents implémentations et configurations avec la collection d'évaluation par analogie en français.

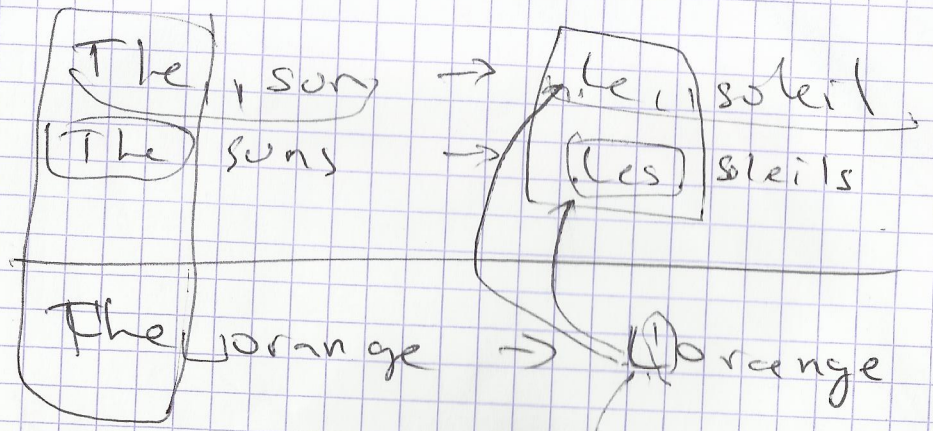
Plusieurs implémentations sont disponibles :

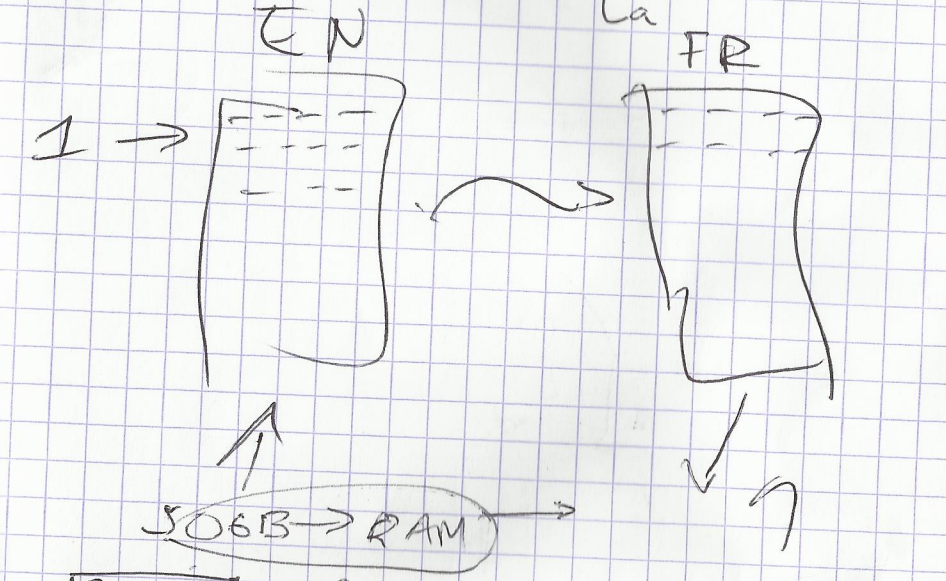
- Gensim https://radimrehurek.com/gensim/
- hyperwords https://bitbucket.org/omerlevy/hyperwords
- word2vec https://code.google.com/archive/p/word2vec/
- tensorflow https://www.tensorflow.org/tutorials/word2vec/

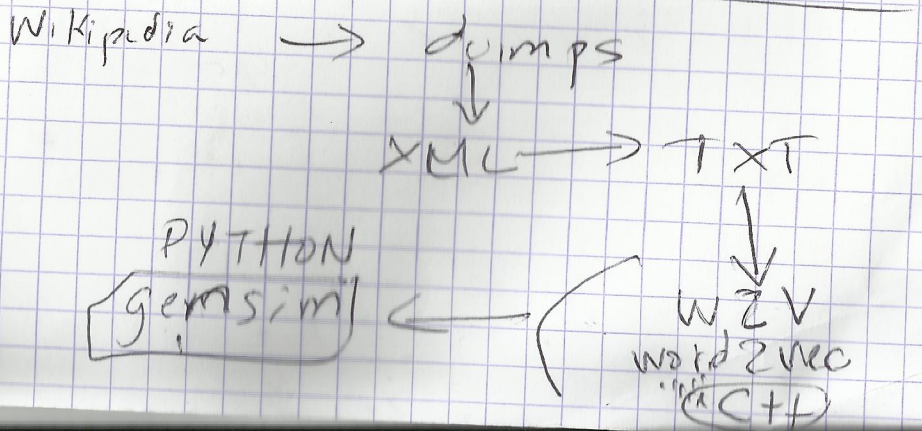
Procédure recommandée :

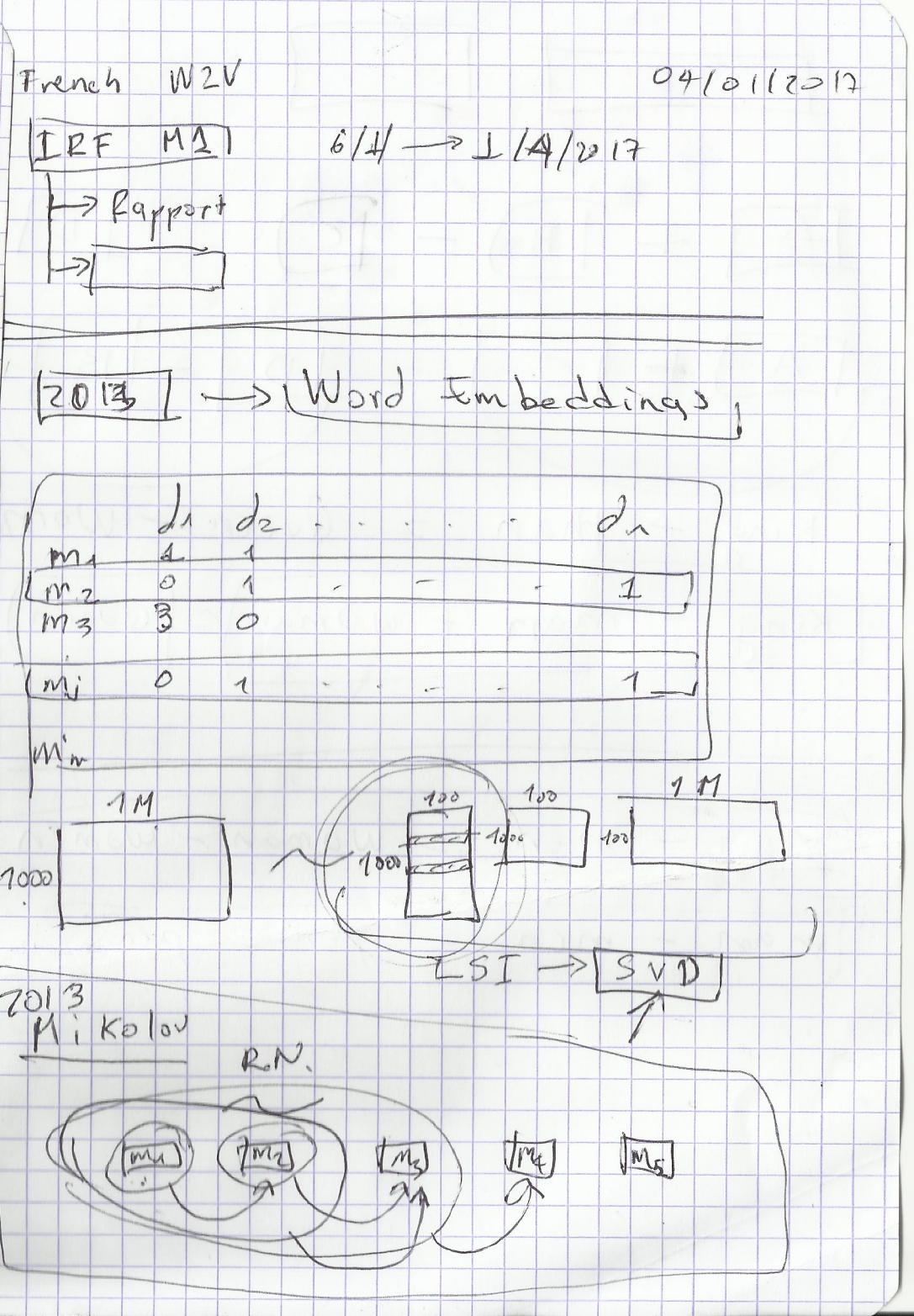
- Télécharger les fichiers questions-words.txt et question-phrases.txt disponibles sur https://code.google.com/archive/p/word2vec/ (chercher autres collections pertinents)
- Trouver les mots uniques dans chaque fichier, les traduire et faire le mapping
- Vérifier que les analogies sont toujours valides (pas de problèmes de traduction) et proposer des analogies équivalents pour le français
- Télécharger la Wikipédia en français (fichier frwiki-20170101-pages-articles.xml.bz2 sur https://dumps.wikimedia.org/frwiki/20170101/ ou le plus récent)
- Traiter la Wikipédia avec les implémentations proposées
- Utiliser les vecteurs obtenus pour calculer la performance avec chaque combinaison des paramètres (au moins deux calcules peut être faites, 3COSADD et 3COSMUL [1])
- Rédiger un rapport avec les résultats et les conclusions de ce projet

[1] Omer Levy and Yoav Goldberg. Linguistic Regularities in Sparse and Explicit Word Representations.
[2] Omer Levy, Yoav Goldberg, Ido Dagan. Improving distributional similarity with lessons learned from word embeddings.
[3] T Mikolov, I Sutskever, K Chen, GS Corrado, J Dean. Distributed representations of words and phrases and their compositionality
[4] T Mikolov, K Chen, G Corrado, J Dean. Efficient estimation of word representations in vector space


The soleil
The suns
les soleils
The orange
orange
La

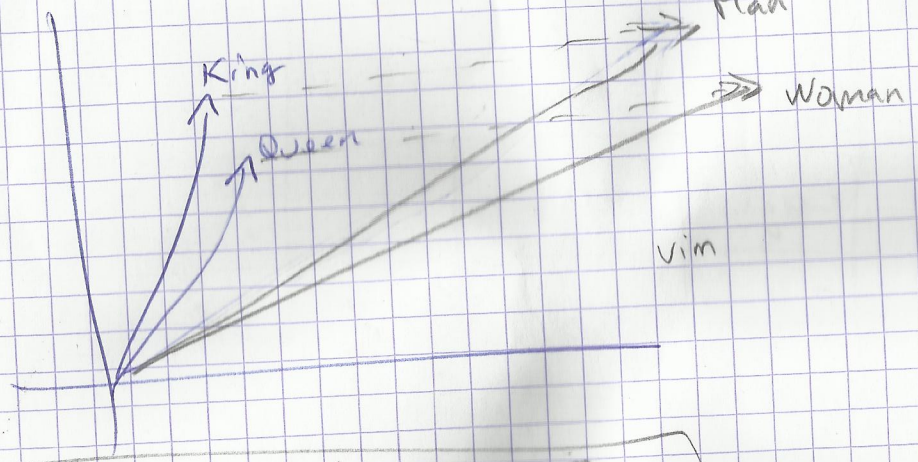

EN
FR
1
SOGB → RAM
SIMPLE → 86


Wikipedia
dumps
XML
TXT
PYTHON
Gensim
W2V
word2vec
C++


French W2V
04/01/2017
IRF M1
6/1 → 1/4/2017
Rapport
2013 → Word Embeddings
d1 d2 dn
m1 1 1
m2 0 1 1
m3 3 0
mi 0 1 1
mn
1M
1000
100
100
1000
1000
100
1 M
LSI → SVD
2013
Mikolov
R.N.
m1
m2
m3
m4
m5

King - Man ≈ Queen - Woman

<Page>
<title> ... </title>



vim

King - Man + Woman ≈ Queen

Le

L'arbre

The / tree

don't

nltk

A + B - C = D

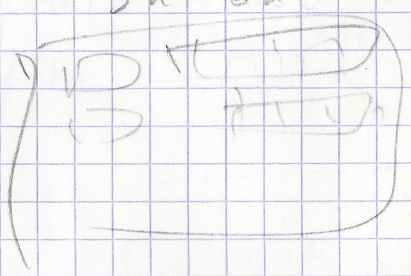
A + C = D - B

King - Man = Queen - Woman

King - man + woman ≈ Queen

man - men = woman - women

man - men = pen - pens



Wipedia

	W = 5		W = 50	
	Skipgram	CBoW	Skipgram	CBoW
English	40%			
French	30%	20%	25%	50%