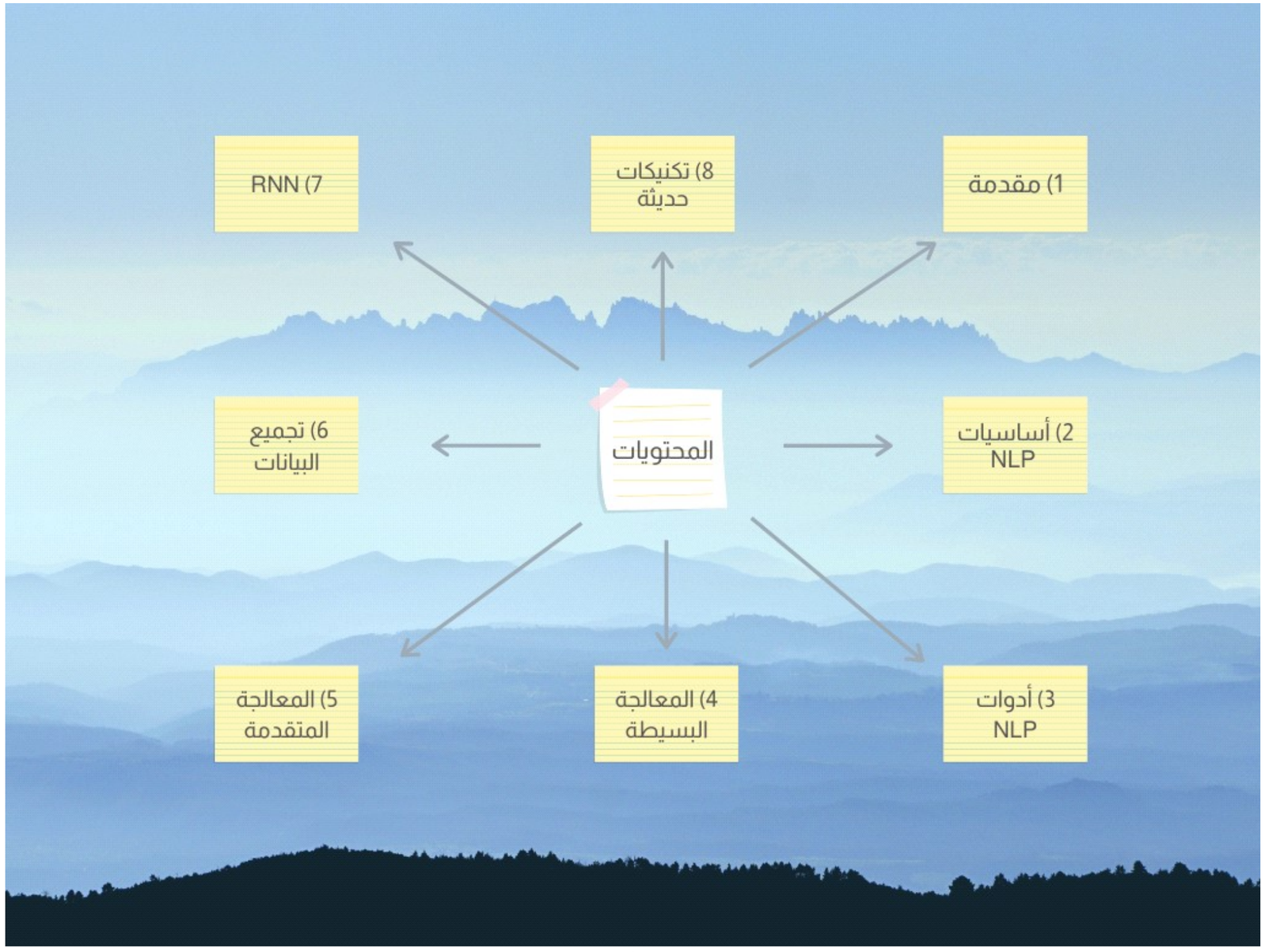
Natural Language Processing المعالجة اللغوية الطبيعية



المحتويات

1) مقدمة	المحتويات	ما هو NLP	تاريخ NLP	العقبات و التحديات	التطبيقات				
2) أساسيات NLP	المكتبات	الملفات النصية	ملفات pdf	البحث في النصوص					
3) أدوات NLP	Tokenization	Sent. Segm.	POS	Stem & Lemm	NER	Stopwords	Matchers	Syntactic Struc.	T.Visualization
4)المعالجة البسيطة	Word Meaning	Word embed	T. Vectors	Word2Vec	BOW	TF-IDF	Text Similarity	Dist. Similarity	
5)المعاجلة المتقدمة	T. Classification	T. Clustering	LDA	NMF	GloVe	Lexicons	NGrams	L. Modeling	T. Generation
	Sent. Analysis	Naïve Bayes	Vader	Auto Correct	Ans. Questions		Summarization & Snippets		
6)تجميع البيانات	Tweet Collecting	Data Scraping	Information Extraction		Information Retrieval		Relative Extraction		Search Engine
RNN (7	Seq to Seq	LSTM	GRU	Rec NN\TNN					
8)تكنيكات حديثة	Word Cloud	CNN	T. Forcing	Attention Model	Transformer	Bert	FastText	Gensim	Chat Bot

القسم الخامس : المعالجة المتقدمة للنصوص

الجزء السادس : GloVe

==========

نتناول الأول اداة بالغة الأهمية هي أداة GloVe , واي اختصار : GloVe: Global Vectors for Word Representation و هي تعني : المصفوفة العامة لإظهار الكلمات

و هي في الأساس خوارزيم تم تدريبه بالفعل علي عدد ضخم من الكلمات , بأسلوب التدريب دون اشراف , للتعرف علي مدي اقتراب الكلمات من بعضها البعض , التعرف علي مصفوفة التضمين للكلمات , ورسم الكلمات القريبة او البعيدة عن بعضها البعض , وهو من ابتكار جامعة ستانفورد , وموجود هنا :

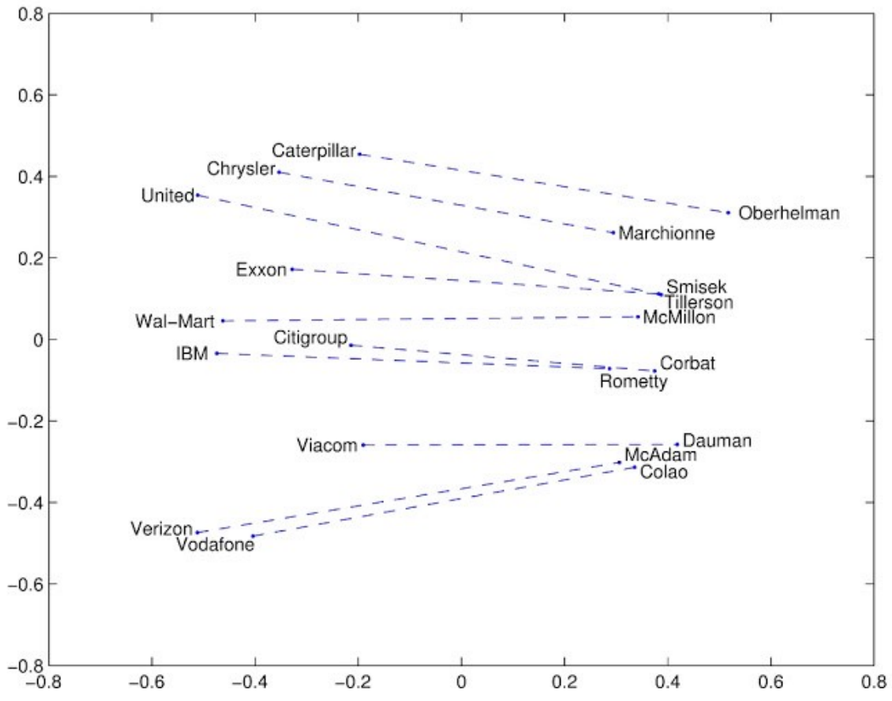
https://github.com/stanfordnlp/GloVe

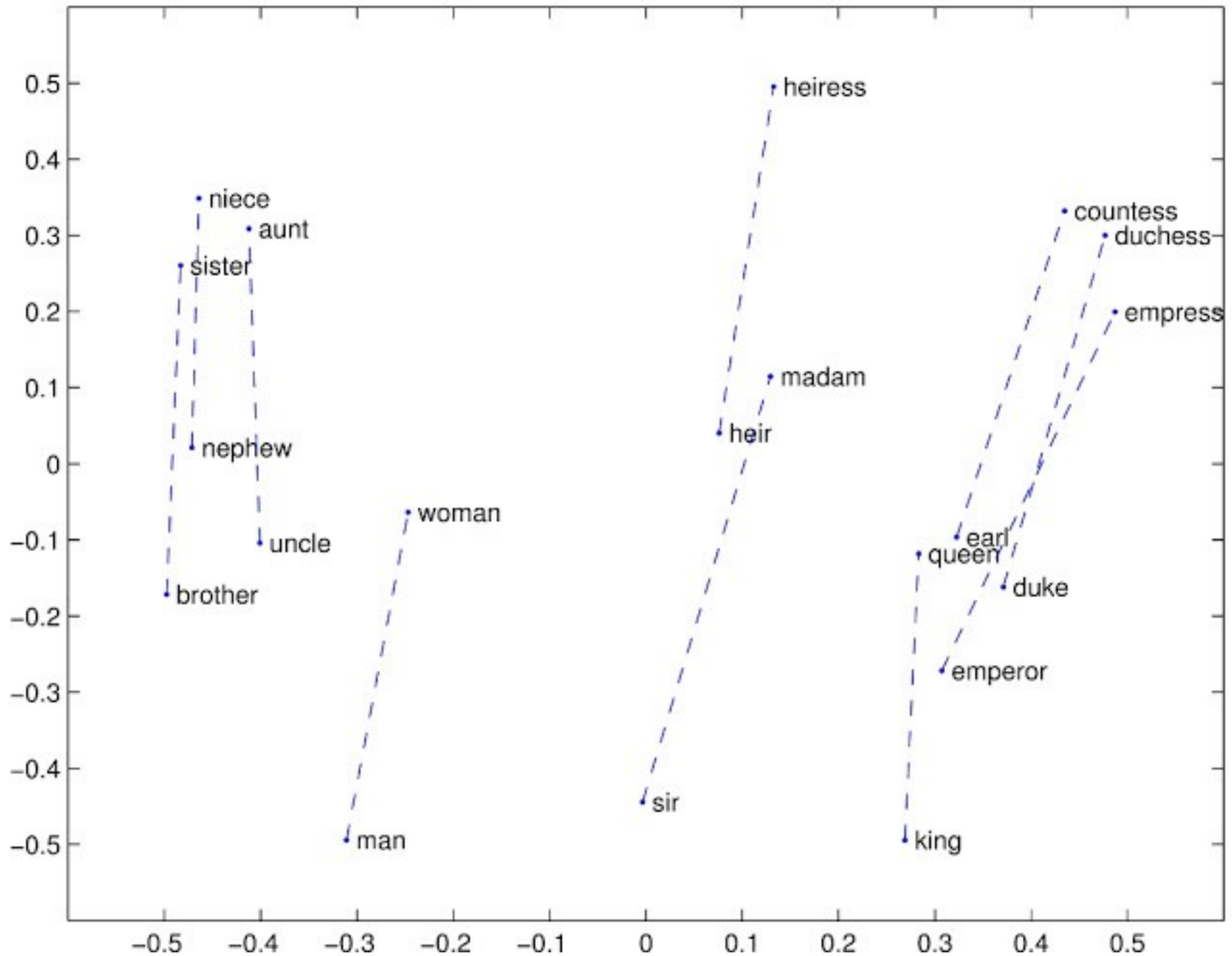
و هنا التفاصيل كاملة :

https://nlp.stanford.edu/projects/glove/

و نقطة التفوق الأساسية للـ GloVe علي Word2Vec و قيم Word Embedding هي أن GloVe لا تقوم بحساب قيم الكلمات علي البيانات التي يتم التدريب عليها , ولكن علي البيانات الشاسعة التي قامت جامعة ستانفورد بتدريبها عليها , وهو ما يجعل الدقة أعلي بكثير , حتي لو كانت الكلمات لديك قليلة

و هي قائمة علي فكرة علاقة الكلمات بعضها البعض




heiress
niece
aunt
countess
duchess
sister
empress
madam
nephew
heir
woman
uncle
earl
queen
brother
duke
emperor
sir
man
king
0.5
0.4
0.3
0.2
0.1
0
−0.1
−0.2
−0.3
−0.4
−0.5
−0.5
−0.4
−0.3
−0.2
−0.1
0
0.1
0.2
0.3
0.4
0.5

_

و تقوم فكرة تدريب GloVe علي حساب ما يسمي : مصفوفة التواجد المشترك Co-Occurrence Matrix

و هي مصفوفة مربعة (الصفوف و الأعمدة متساوي) و يكون فيها جميع الكلمات الموجودة في النص (بعد حذف التكرار) و تقوم القيم المشتركة في الجدول , بتوضيح مدي تواجدها مع الكلمات الأخرى في سياق محدد او في جملة معينة

$$X = \begin{array}{c} \\ I \\ like \\ enjoy \\ deep \\ learning \\ NLP \\ flying \\ . \end{array} \begin{array}{c} \begin{array}{cccccccc} I & like & enjoy & deep & learning & NLP & flying & . \end{array} \\ \begin{bmatrix} 0 & 2 & 1 & 0 & 0 & 0 & 0 & 0 \\ 2 & 0 & 0 & 1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 1 & 1 & 1 & 0 \end{bmatrix} \end{array}$$

و لحساب مصفوفة التواجد المشترك , يجب أن نحدد أولا حجم النافذة , وهي عدد الكلمات التي سيتم النظر اليها يمينا و يسارا , للتعامل معها , فلو كانت 1 , أي أننا سنحسب علاقة كل كلمة , مع الكلمة السابقة لها و التالية لها فحسب , ولو كانت 2 فهما اثنين قبل و اثنين بعد و هكذا , ولكن بشرط ان تكون في نفس الجملة , فلو كانت هناك كلمة في بداية الجملة , فلا ننظر الي الكلمة السابقة لها , حتي لو كانت ملاصقة لها , كذلك الكلمة نهاية الجملة لا ننظر للكلمة التالية لها

فلو كان لدينا جملة بسيطة هي :

I love Programming. I love Math. I tolerate Biology.

و كانت الـ window تساوي 1 , فنجد ان كلمة المصفوفة ستكون :

	I	love	Programming	Math	tolerate	Biology	.
I	0	2	0	0	1	0	2
love	2	0	1	1	0	0	0
Programming	0	1	0	0	0	0	1
Math	0	1	0	0	0	0	1
tolerate	1	0	0	0	0	1	0
Biology	0	0	0	0	1	0	1
.	1	0	1	1	0	1	0

ولو كان لدينا جملة :

Roses are red. Sky is blue

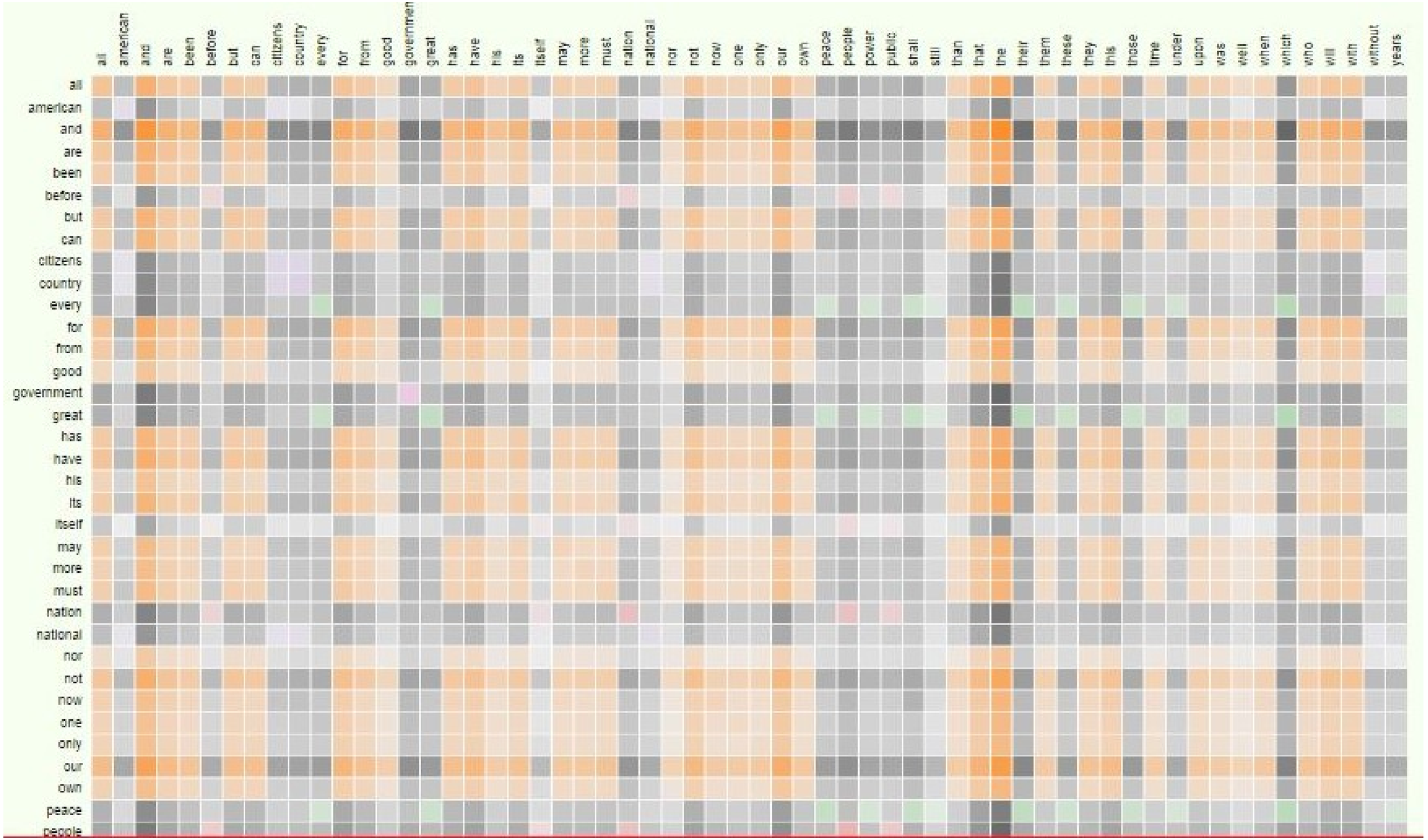
و كانت ال window تساوي 3 , ستكون المصفوفة :

	Roses	are	red	Sky	is	blue
Roses	1	1	1	0	0	0
are	1	1	1	0	0	0
red	1	1	1	0	0	0
Sky	0	0	0	1	1	1
is	0	0	0	1	1	1
Blue	0	0	0	1	1	1

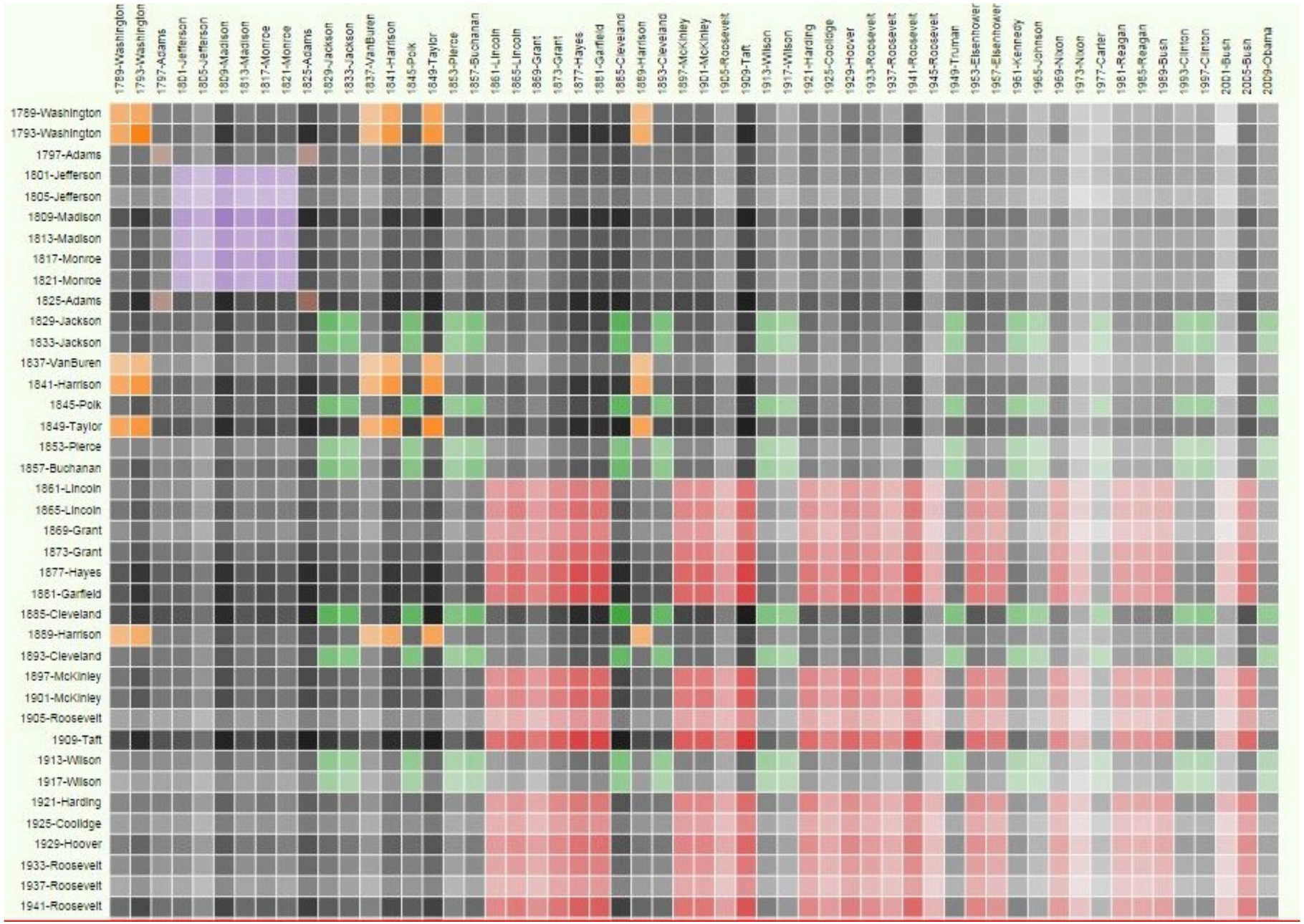
و قد تكون اكبر

counts	I	like	enjoy	deep	learning	NLP	flying	.
I	0	2	1	0	0	0	0	0
like	2	0	0	1	0	1	0	0
enjoy	1	0	0	0	0	0	1	0
deep	0	1	0	0	1	0	0	0
learning	0	0	0	1	0	0	0	1
NLP	0	1	0	0	0	0	0	1
flying	0	0	1	0	0	0	0	1
.	0	0	0	0	1	1	1	0

او اكبر , وبالألوان التي تدل علي الارقام , وهنا مصفوفة عن علاقة الكلمات



وهنا مصفوفة عن علاقة النصوص بعضها البعض



_

و ايضا يجب أن نتعرف علي ما يسمي : جدول نسب الاحتمالات Probability Ratios Table

حيث يمثل كل عمود , الكلمة التي نريد قياسها , و الصف الاول , موضوع عن السفر , والصف الثاني عن المدارس , و تمثل القيمة مدي احتمالية تواجد هذه الكلمة في هذا الموضوع , بينما يمثل الصف الاخير قسمة الصف الاولي علي الثاني

	Passport	Exams	Explore	Money
P(k/Travel)	0.25	0.004	0.26	0.45
P(k/School)	0.003	0.34	0.28	0.41
P(k/Travel) / P(k/School)	83	0.01	0.92	1.09

و زيادة الارقام او تقليلها يشير الي مدي تواجد الكلمة هنا اكثر من هنا

_

و التعامل مع GloVe لا يكون عبر استخدام مكتبة محددة , ولكن بتحميل الملفات التي قامت معامل جامعة ستانفورد بتجهيزها , والتعامل معها بشكل مباشر

هنا جميع الملفات :

https://nlp.stanford.edu/projects/glove/

و سنقوم بتحمل احد الملفات والذي تم تدريبه علي علي 6 مليارات كلمة من هنا :

http://nlp.stanford.edu/data/glove.6B.zip

و نري هنا 4 ملفات glove.6B.50d , glove.6B.100d , glove.6B.200d , glove.6B.300d , يشتمل كل ملف فيهم علي مليارات الكلمات , هي ارقام قيم التضمين embedding الخاصة بكل كلمة , إما 50 او 100او 200 او 300 , هكذا :

```
the 0.418 0.24968 -0.41242 0.1217 0.34527 -0.044457 -0.49688 -0.17862 -0.00066023 -0.6566 0.27843 -0.14767 -0.55677 0.14658 -0.0095095 0.011658 0.10204 -0.12792 -0.8443 -0.12181 -0.016801 -0.33279 -0.1552 -0.23131 -0.19181 -1.8823 -0.76746 0.099051 -0.42125 -0.19526 4.0071 -0.18594 -0.52287 -0.31681 0.00059213 0.0074449 0.17778 -0.15897 0.012041 -0.054223 -0.29871 -0.15749 -0.34758 -0.045637 -0.44251 0.18785 0.0027849 -0.18411 -0.11514 -0.78581
```

, 0.013441 0.23682 -0.16899 0.40951 0.63812 0.47709 -0.42852 -0.55641 -0.364 -0.23938 0.13001 -0.063734 -0.39575 -0.48162 0.23291 0.090201 -0.13324 0.078639 -0.41634 -0.15428 0.10068 0.48891 0.31226 -0.1252 -0.037512 -1.5179 0.12612 -0.02442 -0.042961 -0.28351 3.5416 -0.11956 -0.014533 -0.1499 0.21864 -0.33412 -0.13872 0.31806 0.70358 0.44858 -0.080262 0.63003 0.32111 -0.46765 0.22786 0.36034 -0.37818 -0.56657 0.044691 0.30392

. 0.15164 0.30177 -0.16763 0.17684 0.31719 0.33973 -0.43478 -0.31086 -0.44999 -0.29486 0.16608 0.11963 -0.41328 -0.42353 0.59868 0.28825 -0.11547 -0.041848 -0.67989 -0.25063 0.18472 0.086876 0.46582 0.015035 0.043474 -1.4671 -0.30384 -0.023441 0.30589 -0.21785 3.746 0.0042284 -0.18436 -0.46209 0.098329 -0.11907 0.23919 0.1161 0.41705 0.056763 -6.3681e-05 0.068987 0.087939 -0.10285 -0.13931 0.22314 -0.080803 -0.35652 0.016413 0.10216

of 0.70853 0.57088 -0.4716 0.18048 0.54449 0.72603 0.18157 -0.52393 0.10381 -0.17566 0.078852 -0.36216 -0.11829 -0.83336 0.11917 -0.16605 0.061555 -0.012719 -0.56623 0.013616 0.22851 -0.14396 -0.067549 -0.38157 -0.23698 -1.7037 -0.86692 -0.26704 -0.2589 0.1767 3.8676 -0.1613 -0.13273 -0.68881 0.18444 0.0052464 -0.33874 -0.078956 0.24185 0.36576 -0.34727 0.28483 0.075693 -0.062178 -0.38988 0.22902 -0.21617 -0.22562 -0.093918 -0.80375

to 0.68047 -0.039263 0.30186 -0.17792 0.42962 0.032246 -0.41376 0.13228 -0.29847 -0.085253 0.17118 0.22419 -0.10046 -0.43653 0.33418 0.67846 0.057204 -0.34448 -0.42785 -0.43275 0.55963 0.10032 0.18677 -0.26854 0.037334 -2.0932 0.22171 -0.39868 0.20912 -0.55725 3.8826 0.47466 -0.95658 -0.37788 0.20869 -0.32752 0.12751 0.088359 0.16351 -0.21634 -0.094375 0.018324 0.21048 -0.03088 -0.19722 0.082279 -0.09434 -0.073297 -0.064699 -0.26044

وغالبا ما نقوم بقراءة الملف و وضع القيم في قاموس , ثم عدد من العمليات بين الكلمات المطلوبة مثل :

- عمل مقارنة بين معاني الكلمات
- البحث عن اقرب كلمات لكلمة معينة
- تنفيذ عمليات رياضية في الكلمات king-man+woman

-

كما أن هناك عملية مشابهة لما تم في GloVe و لكن ببيانات مختلفة , و هي بيانات قامت بها جوجل و حجمها يصل الي 3.5 جيجا , وفيها 3 مليون كلمة

كما ان الكلمات هنا بها ميزة , و هي ان الكلمات المرتبطة ببعضها هناك _ بينها , اي مثل : Los_Angeles و هو ما يجعل فهم الكلمات اكثر دقة

لكن هذه الداتا مكتوبة بطريقة binary و ليست كلمات مثل glove , لذا سنستخدم لها مكتبة genism للتعامل معها , و هي ما سنراها في القسم الثامن

-

كما أن هناك داتا ضخمة خاصة باللغة العربية , ويتم تنفيذ نفس المهام عليها , وهي لها 1538616 كلمة و 256 رقم لكل كلمة

و هي هنا بالكامل

https://github.com/tarekeldeeb/GloVe-Arabic

و هنا مثال لها :

word في, with Vecs [1.936900e-01 -6.868660e-01 2.095150e-01 3.070000e-04 1.078010e-01
-1.171430e-01 -1.321890e-01 3.891600e-01 -3.993030e-01 -1.416040e-01
-2.944130e-01 -2.720610e-01 -3.711600e-02 3.048600e-01 1.890560e-01
-1.249200e-02 2.220200e-01 4.954420e-01 1.005328e+00 -2.961000e-03
2.251000e-02 8.873760e-01 -3.452900e-01 -3.586780e-01 -4.460700e-02
3.326800e-02 3.679750e-01 3.888200e-01 -2.443990e-01 7.172520e-01
3.749510e-01 7.283600e-02 2.141690e-01 2.573760e-01 -5.784450e-01
1.729480e-01 1.397000e-01 -2.867760e-01 5.557820e-01 -2.226420e-01
-2.036990e-01 2.897970e-01 2.799230e-01 1.396100e-02 -1.212050e-01
-4.835330e-01 -4.670190e-01 2.949960e-01 2.959800e-02 -3.933150e-01

1.031570e-01 1.275158e+00 -8.975260e-01 -4.661910e-01 9.499000e-02
1.042300e-01 -2.165290e-01 8.959500e-02 -4.467920e-01 -6.975010e-01
-9.327700e-02 -6.182880e-01 5.166490e-01 -9.837000e-02 1.023650e-01
4.371750e-01 -3.085370e-01 5.316160e-01 8.186500e-02 -1.550210e-01
-3.699340e-01 3.749530e-01 3.677500e-02 1.465960e-01 1.173352e+00
1.464860e-01 -1.508550e-01 2.106850e-01 9.188000e-02 -3.571170e-01
-5.103100e-02 4.282170e-01 -2.127300e-01 -1.388930e-01 -3.871140e-01
2.489800e-01 -5.463850e-01 6.980400e-02 1.233310e-01 -4.545830e-01
7.435810e-01 -1.452520e-01 8.463900e-02 -3.108300e-02 3.774570e-01
-6.007700e-02 -3.726000e-03 -1.080123e+00 2.347980e-01 -3.486390e-01
-1.807840e-01 -5.280070e-01 4.826500e-02 4.247610e-01 -8.514710e-01
-9.617800e-02 4.290630e-01 1.349600e-01 -7.241100e-02 9.685800e-02
8.601700e-02 1.470070e-01 3.365410e-01 -7.711000e-03 -2.452400e-01
3.381370e-01 -1.972030e-01 -5.027980e-01 -1.757610e-01 -9.700460e-01
3.273600e-01 -2.505730e-01 -7.400000e-05 -1.176551e+00 5.860400e-02
1.956260e-01 -4.106400e-01 -3.144980e-01 -1.895260e-01 -4.194750e-01
4.667590e-01 -2.059970e-01 -2.933560e-01 9.278000e-03 2.950200e-02
-1.225010e-01 -2.737470e-01 -1.117750e-01 -2.721660e-01 -4.489860e-01
-5.125800e-02 5.275700e-02 -6.905800e-02 -6.128600e-02 2.177730e-01
-1.491830e-01 -9.133200e-02 2.401150e-01 -3.917100e-02 -3.964000e-02
5.257910e-01 -1.742030e-01 1.154330e-01 -4.854380e-01 4.510920e-01

-1.254660e-01 4.629150e-01 7.297420e-01 -5.992093e+00 -1.698200e-01
 1.826260e-01 2.876010e-01 3.779250e-01 6.002670e-01 4.044390e-01
 6.584050e-01 -4.959260e-01 1.207950e-01 6.776000e-02 5.299160e-01
-4.510680e-01 -5.066190e-01 6.628800e-02 1.746930e-01 2.667520e-01
 4.882110e-01 -8.330400e-02 -1.604180e-01 9.491800e-02 -7.938700e-02
-1.183520e-01 2.907320e-01 7.466600e-02 2.952300e-02 -2.132000e-02
 9.404500e-02 -1.637200e-01 -1.711080e-01 7.753400e-02 2.768950e-01
 2.329650e-01 -6.171900e-02 -4.344390e-01 -2.081640e-01 -7.588230e-01
 1.527900e-01 5.592350e-01 -1.671760e-01 -1.995410e-01 3.811590e-01
 2.386560e-01 -9.298000e-03 -2.801000e-02 4.295580e-01 -3.539750e-01
 4.455340e-01 2.641530e-01 -4.253360e-01 -5.029400e-02 -6.291040e-01
-4.072260e-01 5.795500e-02 -2.080010e-01 -2.462200e-02 4.182650e-01
-1.627820e-01 2.483400e-02 -2.872310e-01 1.469210e-01 1.846030e-01
-1.496140e-01 -1.305170e-01 1.811820e-01 4.171700e-01 -3.190300e-02
 3.125530e-01 -2.276730e-01 1.497870e-01 3.090700e-01 3.680610e-01
 5.445190e-01 -1.706940e-01 -1.335720e-01 -1.815760e-01 -3.690700e-02
-5.170000e-02 -3.203660e-01 -4.463370e-01 -1.407700e-02 -3.359150e-01
 1.172580e-01 -3.125170e-01 6.932100e-02 1.871700e-02 -1.016210e-01
-1.812890e-01 -1.259010e-01 -4.452880e-01 -1.971170e-01 3.250490e-01
-2.623990e-01 2.690800e-02 1.163850e-01 -3.147030e-01 6.897560e-01
 4.169670e-01]

word من, with Vecs [2.644500e-01 -1.149580e-01 -1.762000e-01 -3.219870e-01 5.510360e-01
3.176180e-01 4.485640e-01 2.036210e-01 -2.887030e-01 -7.160000e-03
1.583810e-01 4.577900e-02 -1.241740e-01 -1.887400e-02 5.031500e-02
-2.503020e-01 -3.283400e-02 -2.238250e-01 2.687010e-01 -3.546450e-01
5.587290e-01 3.645380e-01 1.850990e-01 3.678700e-02 -1.906800e-01
3.384210e-01 -7.296500e-02 1.880950e-01 -3.366440e-01 2.890510e-01
2.121600e-02 1.033020e-01 1.972100e-01 6.582880e-01 -1.111500e-01
-2.301800e-02 -1.806450e-01 -3.562380e-01 4.295600e-02 -4.280200e-02
-1.045850e-01 4.985970e-01 -2.973530e-01 -5.379580e-01 1.439840e-01
3.644440e-01 -7.538700e-02 -2.131630e-01 -4.443700e-02 -6.214390e-01
1.628700e-01 1.042009e+00 2.597000e-03 2.929510e-01 3.058820e-01
1.200540e-01 -2.838150e-01 2.609200e-02 1.299720e-01 1.451780e-01
1.180300e-01 3.924080e-01 -3.770820e-01 3.668980e-01 -4.704300e-02
-1.224380e-01 -1.275540e-01 -5.287180e-01 -4.065410e-01 1.975660e-01
-6.046220e-01 -3.474630e-01 -2.416370e-01 -7.112420e-01 5.652140e-01
-2.949860e-01 -1.463030e-01 4.020590e-01 -1.742070e-01 -1.090190e-01
-1.779620e-01 -2.779510e-01 6.089890e-01 2.552990e-01 -4.911300e-02
-1.193800e-02 3.425200e-02

_