Knowledge extraction from large text and source code multimodal document collections using machine learning

Ing. Juan Felipe Baquero
Grupo MindLab
Universidad Nacional de Colombia

Contenido

1. Introducción
2. Problema
3. Modelo - ¿Qué estoy haciendo?
    1. DataSet
4. Word2vec
    1. ¿Por qué word2vec?
    2. Uno propio vs uno general
    3. Entrenamiento
    4. Resultados

Introducción

¿Cuando tiene un problema que hacen?

... ¿Problemas de programador?
EJ: El codigo no Compila

Introducción

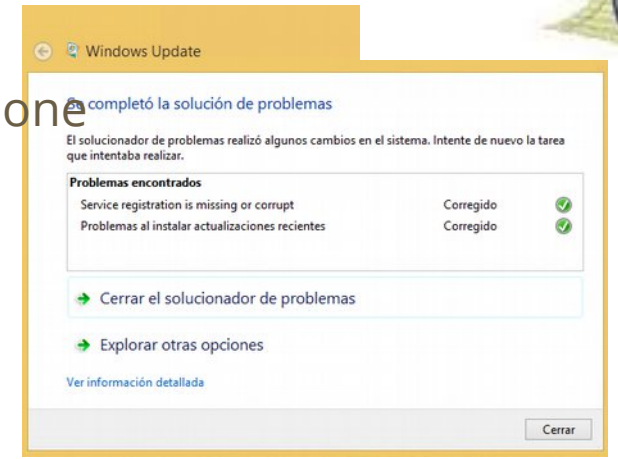
¿Cuando tiene un problema que hacen?

... ¿Problemas de programador?
EJ: El código no Compila/Ejecuta

A) Llamar a un amigo.

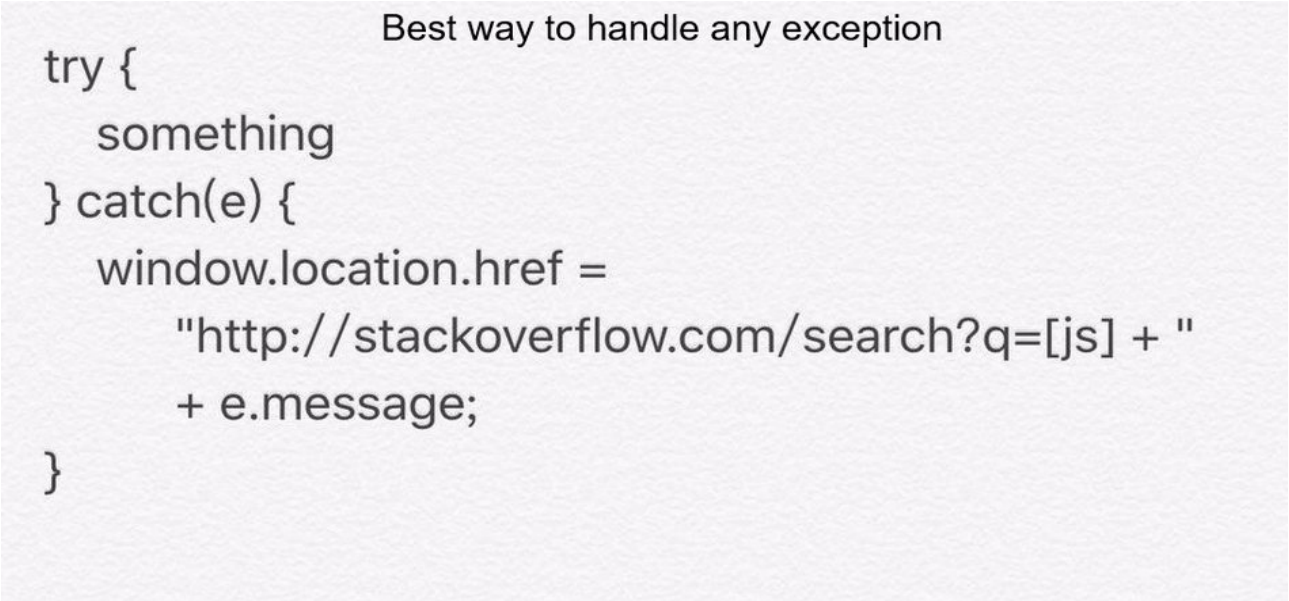
B) Rezar y oprimir F5 hasta que funcione

C) O ...



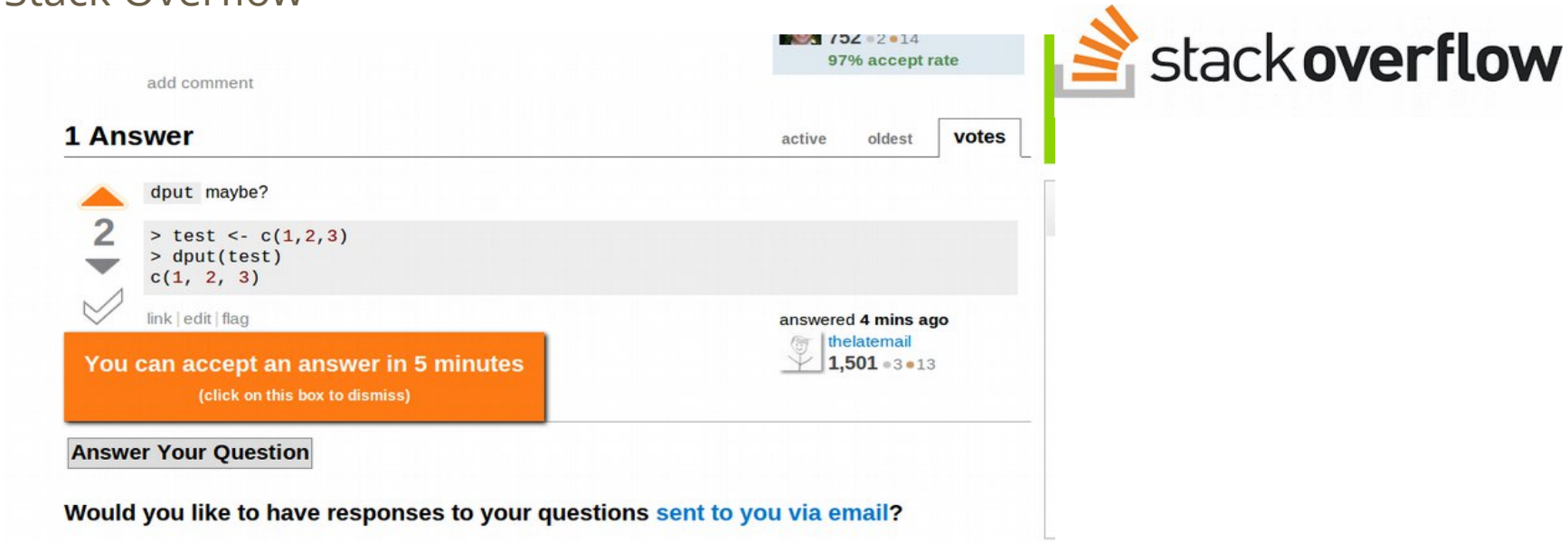
Introducción

Stack Overflow



Introducción

Stack Overflow



Problema

Repositorios de código = Mucha información

¿Se puede obtener información valiosa?

Versionamiento de código
Comunicación entre personas
Tipos de proyectos
Seguimiento de defectos, etc.

¿Cómo se puede obtener esa información?

Problema - Justificación

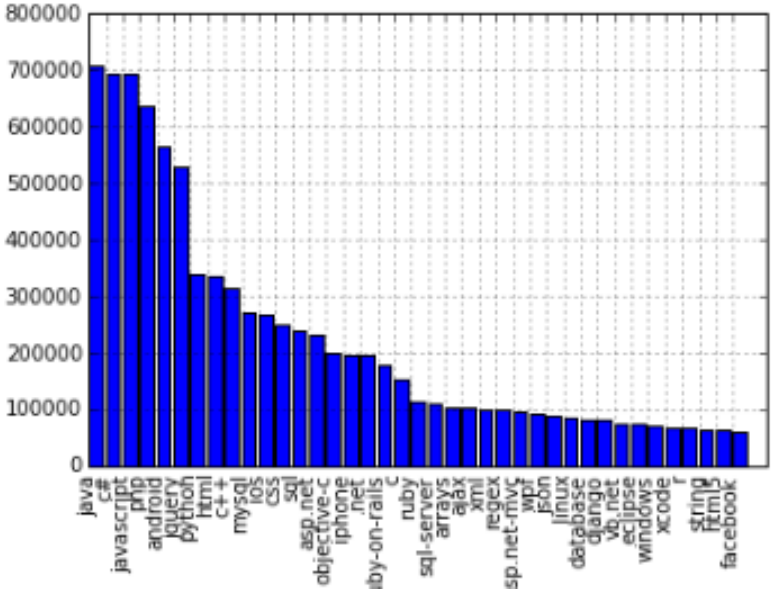
Poder explotar y entender esta valiosa información aplicándola sobre sistemas y proyectos de software, permite la mejorar, control del progreso, mantenimiento y evolución del software entre otros.

Datos

Cantidad por Tipos de Post

	Question	Answer	Orphaned tag wiki	Tag wiki excerpt	Tag wiki	Moderator nominatio	Wiki placeholder	Privilege wiki
0	7990787	13684117	167	30659	30658	200	4	2

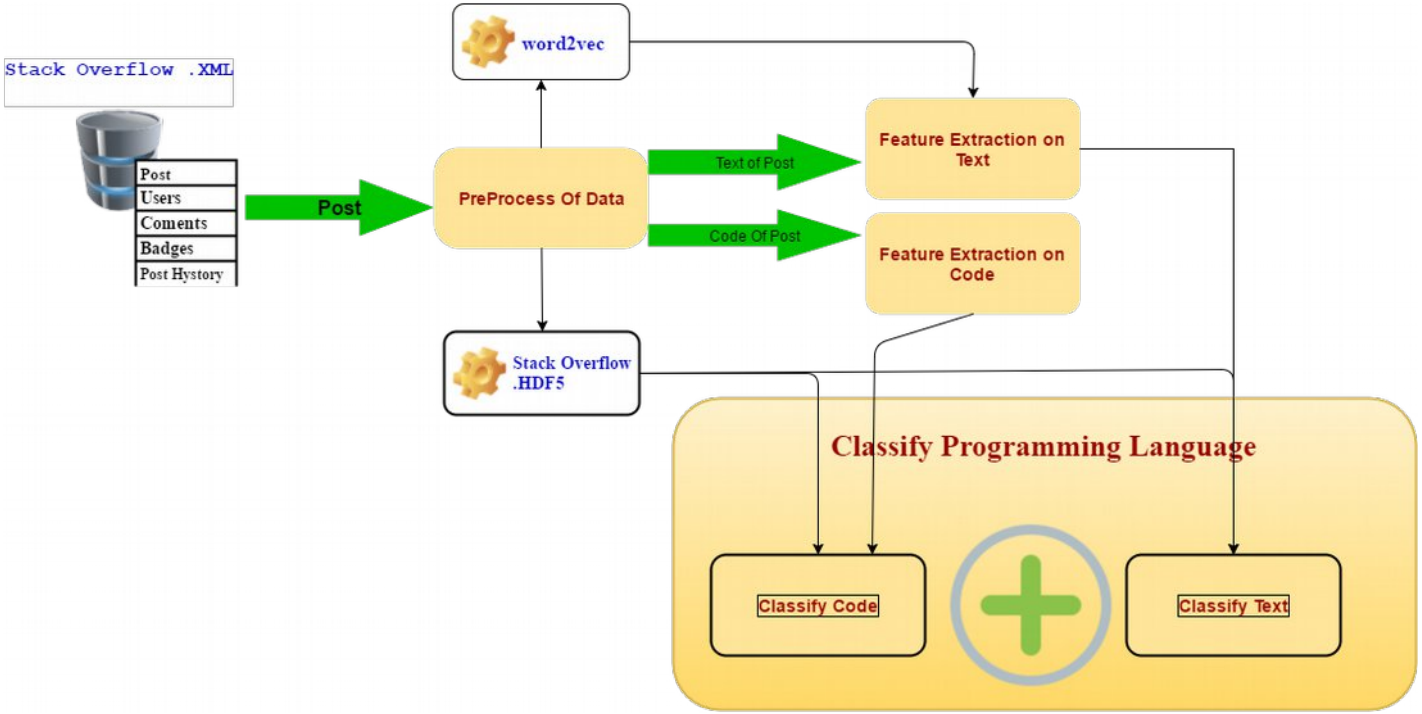
```
Parent Directory                                    -
stackoverflow.com-Ta..> 2014-10-07 07:35  508K
stackoverflow.com-Po..> 2014-10-07 08:41   26M
stackoverflow.com-Ba..> 2014-10-08 08:36   78M
stackoverflow.com-Us..> 2014-10-07 08:45  101M
stackoverflow.com-Vo..> 2014-10-07 08:58  385M
stackoverflow.com-Co..> 2014-10-07 08:40  1.8G
stackoverflow.com-Po..> 2014-10-08 06:29  5.7G
stackoverflow.com-Po..> 2014-10-08 08:31  9.4G
46G PostHistory.xml   8.0G Comments.xml   926M Badges.xml   226M PostLink
29G Posts.xml         6.1G Votes.xml      895M Users.xml    3.2M Tags.xml
```



Numero de TAGS: 38206

Datos MSR2015
publicados 2014-10-08

Modelo


Stack Overflow .XML
Post
Users
Coments
Badges
Post Hystory
Post
word2vec
PreProcess Of Data
Text of Post
Code Of Post
Feature Extraction on Text
Feature Extraction on Code
Stack Overflow .HDF5
Classify Programming Language
Classify Code
Classify Text

Uno propio vs uno general

Usar un word2vec general(googleNews) o crear uno nuevo

- Contexto-Uso de las palabras
- Diferente vocabulario, Lenguaje/Nuevas palabras

Uno propio vs uno general
Trabajo Previo w2v

Profane DataSet

Número de Oraciones: 1488
Tamaño del Vocabulario: 2374

Biblia, bible-kjv

Número de Oraciones: 30103
Tamaño del Vocabulario: 13769

Uno propio vs uno general
Trabajo Previo w2v

Profane DataSet

- .. why did that commit ever even get far enough to get to me? ... Either way, it shows a rather distinct lack of actual testing, wouldn't you say? I really see no excuse for crap like this. ...Linus "not happy" Torvalds
- What the heck is your problem? Go back and read it. If it wasn't loaded before, THEN IT WASN'T WORKING BEFORE EITHER! ... Why the hell do you keep on harping on idiotic issues? Stop being a moron, just repeat after me: A caching firmware loader fixes all these issues and is simple to boot. Stop the idiotic blathering already.

Biblia, bible-kjv

1 He who dwells in the shelter of the Most High

will abide in the shadow of the Almighty.

2 I will say to the Lord, "My refuge and my fortress,

my God, in whom I trust."

3 For he will deliver you from the snare of the fowler

and from the deadly pestilence.

Salmos 91

Uno propio vs uno general
Trabajo Previo w2v

Profane DataSet

```
model2.most_similar(positive=['Torvalds'])
```

```
[(u'since', 0.20154866576194763),
 (u'notion', 0.2011907994747162),
 (u'advantage', 0.19940011203289032),
 (u'seriously', 0.19606056809425354),
 (u'liner', 0.19354811310768127),
 (u'claim', 0.19196699559688568),
 (u'ONLY', 0.191707581281662),
 (u'Just', 0.19120000302791595),
 (u'ABSOLUTELY', 0.1909940391778946),
 (u'merges', 0.18669575452804565)]
```

Biblia, bible-kjv

```
model.most_similar(positive=['God'])
```

```
[(u'faith', 0.5748561024665833),
 (u'Christ', 0.5527634024620056),
 (u'Father', 0.5336076617240906),
 (u'Jesus', 0.5165109634399414),
 (u'mystery', 0.5018347501754761),
 (u'obedience', 0.5007921457290649),
 (u'heavenly', 0.49376559257507324),
 (u'sufferings', 0.4920486509799957),
 (u'also', 0.47291165590286255),
 (u'success', 0.46738284826278687)]
```

Uno propio vs uno general
Trabajo Previo w2v

Profane DataSet

```
model2.most_similar(positive=['Torvalds', 'happy'])
```

```
[(u'WRONG', 0.16415317356586456),
 (u'cleanup', 0.15843060612678528),
 (u'description', 0.15359379351139069),
 (u'control', 0.15221193432807922),
 (u'uses', 0.14432014524936676),
 (u'IDENTICALLY', 0.14153727889060974),
 (u'Congratulations', 0.14031358063220978),
 (u'grumpy', 0.13889750838279724),
 (u'worth', 0.13583414256572723),
 (u'proof', 0.13455604016780853)]
```

Biblia, bible-kjv

```
model.most_similar(positive=['God', 'death'])
```

```
[(u'righteousness', 0.5999968647956848),
 (u'life', 0.5875349640846252),
 (u'obedience', 0.5753971338272095),
 (u'iniquity', 0.5669256448745728),
 (u'judgment', 0.5620095729827881),
 (u'understanding', 0.5595765709877014),
 (u'faith', 0.558514416217804),
 (u'repentance', 0.5375586152076721),
 (u'fools', 0.5300117135047913),
 (u'strength', 0.5052989721298218)]
```

Uno propio vs uno general
Trabajo Previo w2v

Profane DataSet

```
model2.most_similar(positive=['please', 'wrong', 'code'])
```

```
[(u'don', 0.9569776058197021),
 (u'I', 0.954272985458374),
 (u'not', 0.9474494457244873),
 (u'f*ck', 0.9471840858459473),
 (u'so', 0.9465514421463013),
 (u'stupid', 0.9429414272308350),
 (u'think', 0.9429131150245667),
 (u'it', 0.9425508379936218),
 (u'moron', 0.9417970180511475),
 (u'understand', 0.9403474926948547)]
```

```
model2.most_similar(positive=['kernel', 'linus'])
```

```
[(u'fart', 0.2176447957754135),
 (u'Instead', 0.20319823920726776),
 (u'wasn', 0.19949883222579956),
 (u'intermixing', 0.19716107845306396),
 (u'bit', 0.19572293758392334),
 (u'top', 0.19480329751968384),
 (u'time', 0.1929250955581665),
 (u'play', 0.18816283345222473),
 (u'problems', 0.1858050525188446),
 (u'definition', 0.185082346200943)]
```

```
model2.most_similar(positive=['kernel'])
```

```
[(u'questions', 0.26729869842529297),
 (u'hurt', 0.2433892786502838),
 (u'Rik', 0.2256496548652649),
 (u'intermixing', 0.22434201836585999),
 (u'wasn', 0.22361327707767487),
 (u'match', 0.22153455018997192),
 (u'surely', 0.21886643767356873),
 (u'explicit', 0.2170693576335907),
 (u'problems', 0.2160271406173706),
 (u'ONLY', 0.2125125527381897)]
```

Uno propio vs uno general
Trabajo Previo w2v

Profane DataSet

```
model2.most_similar(positive=['hope'])
```

```
[(u'insanity', 0.5223914980888367),
 (u'get', 0.5146161317825317),
 (u'make', 0.5133464336395264),
 (u'makes', 0.5114476680755615),
 (u'things', 0.5098432302474976),
 (u'-', 0.5096161365509033),
 (u'sane', 0.5074836015701294),
 (u'and', 0.5066993236541748),
 (u'pull', 0.5059151649475098),
 (u',', 0.5056688785552979)]
```

Biblia, bible-kjv

```
model.most_similar(positive=['hope'])
```

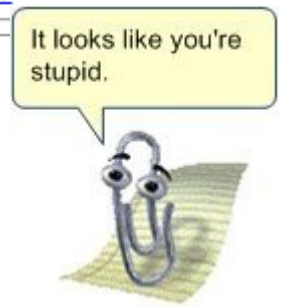
```
[(u'soundness', 0.7748486399650574),
 (u'salvation', 0.7733190059661865),
 (u'favour', 0.7634747624397278),
 (u'delight', 0.7614162564277649),
 (u'strength', 0.7560969591140747),
 (u'secret', 0.735589325428009),
 (u'truth', 0.7212886810302734),
 (u'manifold', 0.7093289494514465),
 (u'folly', 0.703535258769989),
 (u'helper', 0.70188307762146)]
```

Trabajo Previo w2v

¿Que pasaria si se hiciera un ayudante como Clippy el cual fuera entrenado con los correos de Linus Torvalds?

By afcruzs, contest: Codeforces Round #130 (Div. 2), problem: (A) Dubstep, Wrong answer on test 2, #

```
import java.util.*;

public class Dubstep {

        public static void main (String[] args){
                Scanner cin = new Scanner(System.in);

                String dubstep = cin.next();
                String[] original = dubstep.split("WUB");


                System.out.print(original[1]);
                for (int i = 0; i<original.length; i++ ) {
                        if( !original[i].equals(" ") ) System.out.print(original[i] + " ");
                }

        }
}
```



Uno propio vs uno general

Usar un word2vec general(googleNews) o crear uno nuevo

- Contexto-Uso de las palabras
- Diferente vocabulario, Lenguaje/Nuevas palabras
- Publico Objetivo
- Extender el modelo

Uno propio vs uno general

Stack Overflow

```
w2vSO3M.most_similar(positive=['Java'])
```

```
[(u'JAVA', 0.7947582006454468),
 (u'Scala', 0.7388566732406616),
 (u'Python', 0.7271060943603516),
 (u'Perl', 0.683772087097168),
 (u'Ruby', 0.6599066853523254),
 (u'java', 0.6540440917015076),
 (u'Clojure', 0.6473405361175537),
 (u'Groovy', 0.6422683596611023),
 (u'PHP', 0.6398001909255981),
 (u'Haskell', 0.6275602579116821)]
```

GoogleNews

```
w2vGoogle.most_similar(positive=['Java'])
```

```
[(u'Governor_Ahmad_Heryawan', 0.6196017265319824),
 (u'Garut_regency', 0.616998016834259),
 (u'Jumpin_Juice', 0.6157107949256897),
 (u'Surakarta_Central', 0.6124047636985779),
 (u'Pasuruan_East', 0.606446385383606),
 (u'Sukoharjo_Central', 0.60179603099823),
 (u'J2EE', 0.6002236604690552),
 (u'specification_JSR', 0.5977612733840942),
 (u'Tulungagung', 0.5945775508880615),
 (u'Data_Objects_JDO', 0.5928086638450623)]
```

Uno propio vs uno general

Stack Overflow

```
w2vSO3M.most_similar(positive=['IDE'])
```

```
[(u'ide', 0.6373751163482666),
 (u'Eclipse', 0.6337193250656128),
 (u'debugger', 0.6307451128959656),
 (u'CDT', 0.614564061164856),
 (u'editor', 0.6041334271430969),
 (u'interpreter', 0.598480224609375),
 (u'NetBeans', 0.5945308208465576),
 (u'Netbeans', 0.5905935764312744),
 (u'PyDev', 0.5892724990844727),
 (u'compiler', 0.5818835496902466)]
```

GoogleNews

```
w2vGoogle.most_similar(positive=['IDE'])
```

```
[(u'debugger', 0.6048680543899536),
 (u'KDevelop', 0.5807616710662842),
 (u'Visual_SlickEdit', 0.5785201191902161),
 (u'Eclipse_IDE', 0.5774178504943848),
 (u'Environment_IDE', 0.5764614343643188),
 (u'IDEs', 0.5762472152709961),
 (u'3rdRail', 0.5747032165527344),
 (u'GCC_compiler', 0.5739843845367432),
 (u'IntelliJ', 0.5737760066986084),
 (u'IAR_Embedded_Workbench', 0.5716899633407593)]
```

Uno propio vs uno general

```
w2vGoogle.most_similar(positive=['IDE', 'python'])
```

```
[(u'Python', 0.5482786893844604),
 (u'compiler_debugger', 0.5352429151535034),
 (u'VS.NET', 0.517549991607666),
 (u'servlet_container', 0.514489471912384),
 (u'scripting_languages', 0.5120866298675537),
 (u'debugger', 0.5069085955619812),
 (u'pythons', 0.5030524134635925),
 (u'Burmese_python', 0.4962586462497711),
 (u'TITLE_Debian_update', 0.4930167496204376),
 (u'Mono_runtime', 0.49293217062950134)]
```

```
w2vSO3M.most_similar(positive=['IDE', 'python'])
```

```
[(u'Python', 0.7550843358039856),
 (u'netbeans', 0.6844033002853394),
 (u'Eclipse', 0.6826522350311279),
 (u'Perl', 0.6804254055023193),
 (u'eclipse', 0.6732154488563538),
 (u'Netbeans', 0.6707097887992859),
 (u'Jython', 0.6698298454284668),
 (u'IronPython', 0.6694133877754211),
 (u'PyDev', 0.6688504219055176),
 (u'jython', 0.6674999594688416)]
```

Uno propio vs uno general

¿Mejor IDE para Java?

```
w2vGoogle.most_similar(positive=['IDE', 'Java', 'best'])[0]
```

```
(u'IntelliJ', 0.6529332399368286)
```

```
w2vSO3M.most_similar(positive=['IDE', 'Java', 'best'])[0]
```

```
(u'Eclipse', 0.6181491613388062)
```

Uno propio vs uno general

Stack Overflow

```
w2vSO3M.most_similar(positive=['IDE', 'Java'])
```

```
[(u'Eclipse', 0.7350654006004333),
 (u'NetBeans', 0.7170696258544922),
 (u'Netbeans', 0.713869035243988),
 (u'JAVA', 0.6870511770248413),
 (u'Scala', 0.6819231510162354),
 (u'Groovy', 0.6527713537216187),
 (u'Clojure', 0.6481128334999084),
 (u'Python', 0.6470000147819519),
 (u'netbeans', 0.6350607872009277),
 (u'GUI', 0.6178698539733887)]
```

GoogleNews

```
w2vGoogle.most_similar(positive=['IDE', 'Java'])
```

```
[(u'IntelliJ', 0.6725164651870728),
 (u'scripting_languages', 0.6613301038742065),
 (u'J2EE', 0.6588980555534363),
 (u'Eclipse_IDE', 0.6574351787567139),
 (u'Java_IDE', 0.6537434458732605),
 (u'Eclipse_RCP', 0.6507059335708618),
 (u'specification_JSR', 0.6445636749267578),
 (u'PHP', 0.6435267925262451),
 (u'Netbeans', 0.6415724754333496),
 (u'IDEs', 0.6396722793579102)]
```

Uno propio vs uno general

Mas similar entre Java y Python

Mas similar entre Apple y Phone

Uno propio vs uno general

```
print w2vGoogle.most_similar(positive=['eclipse', 'python'])
```

```
[(u'snake', 0.52595055103302), (u'pythons', 0.5137718915939331), (u'reticulated_python', 0.5113823413848877), (u'Burmese_pytho
n', 0.5029097199440002), (u'annular_eclipse', 0.4914274215698242), (u'crocodile', 0.49038201570510864), (u'lizard', 0.4853903949
2607117), (u'reptile', 0.4832247793674469), (u'monitor_lizard', 0.4758228063583374), (u'king_cobra', 0.4690660834312439)]
```

```
print w2vSO3M.most_similar(positive=['eclipse', 'python'])
```

```
[(u'netbeans', 0.803745448589325), (u'Eclipse', 0.7666647434234619), (u'Python', 0.752955436706543), (u'Netbeans', 0.7136954069
137573), (u'jython', 0.7072967290878296), (u'NetBeans', 0.6958203911781311), (u'vim', 0.6934797763824463), (u'cygwin', 0.675965
3091430664), (u'emacs', 0.6479862928390503), (u'Vim', 0.6474020481109619)]
```

```
print w2vGoogle.most_similar(positive=['apple', 'phone'])
```

```
[(u'cellphone', 0.6108524799346924), (u'cell_phone', 0.6088621020317078), (u'telephone', 0.5787233114242554), (u'blackberry', 0.
5512832999229431), (u'handset', 0.5357351303100586), (u'apples', 0.5351675748825073), (u'phones', 0.5341318845748901), (u'Phon
e', 0.5301409363746643), (u'blackberries', 0.519190788269043), (u'pear', 0.5167704224586487)]
```

```
print w2vSO3M.most_similar(positive=['apple', 'phone'])
```

```
[(u'iphone', 0.6235973238945007), (u'cellphone', 0.6151981353759766), (u'handset', 0.6056375503540039), (u'device', 0.5924206376
075745), (u'iPhone', 0.590326189994812), (u'ipad', 0.5874110460281372), (u'Apple', 0.5830206274986267), (u'carrier', 0.582233846
1875916), (u'itunes', 0.582039475440979), (u'AppStore', 0.5807474851608276)]
```

Uno propio vs uno general

```
print w2vGoogle.most_similar(positive=['apple', 'phone'])
```

```
[(u'cellphone', 0.6108524799346924), (u'cell_phone', 0.6088621020317078), (u'telephone', 0.5787233114242554), (u'blackberry', 0.5512832999229431), (u'handset', 0.5357351303100586), (u'apples', 0.5351675748825073), (u'phones', 0.5341318845748901), (u'Phone', 0.5301409363746643), (u'blackberries', 0.519190788269043), (u'pear', 0.5167704224586487)]
```

```
print w2vSO3M.most_similar(positive=['apple', 'phone'])
```

```
[(u'iphone', 0.6235973238945007), (u'cellphone', 0.6151981353759766), (u'handset', 0.6056375503540039), (u'device', 0.5924206376075745), (u'iPhone', 0.590326189994812), (u'ipad', 0.5874110460281372), (u'Apple', 0.5830206274986267), (u'carrier', 0.5822338461875916), (u'itunes', 0.582039475440979), (u'AppStore', 0.5807474851608276)]
```

```
print w2vGoogle.most_similar(positive=['Apple', 'phone'])
```

```
[(u'iPhone', 0.694725751876831), (u'phones', 0.6576014757156372), (u'handset', 0.6574603915214539), (u'cell_phone', 0.6459723711013794), (u'smartphone', 0.6437394022941589), (u'cellphone', 0.6340402960777283), (u'iPhones', 0.6311697959899902), (u'Nexus_One', 0.6189281344413757), (u'Palm_Pre', 0.6094273328781128), (u'Phone', 0.6077129244804382)]
```

```
print w2vSO3M.most_similar(positive=['Apple', 'phone'])
```

```
[(u'iPhone', 0.6759551167488098), (u'device', 0.6485309600830078), (u'Nokia', 0.6148876547813416), (u'iphone', 0.597946286201477), (u'iPad', 0.5972129106521606), (u'iTunes', 0.5854545831680298), (u'smartphone', 0.5730289816856384), (u'BlackBerry', 0.5672226548194885), (u'HTC', 0.5639272928237915), (u'emulator', 0.5635538101196289)]
```

Uno propio vs uno general

¿Framework WEB Para Java?

```
print w2vGoogle.most_similar(positive=['java', 'Django'], negative=['python'])
```

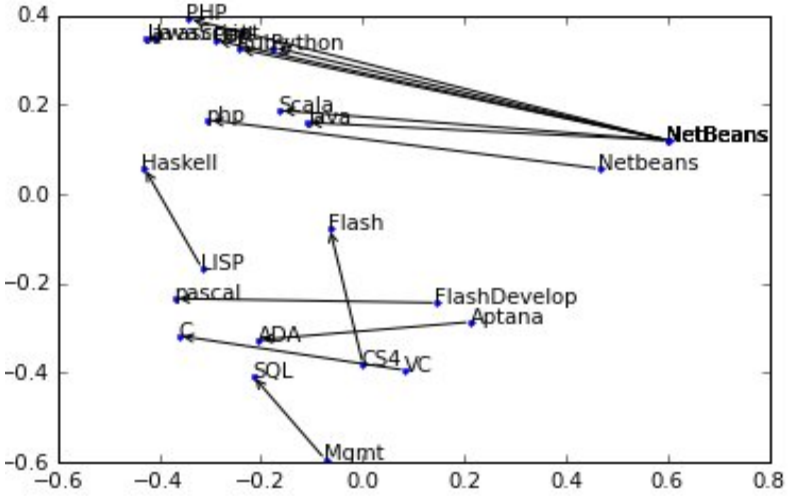
```
[(u'Stumptown_coffee', 0.4839845895767212), (u'o_joe', 0.4637148380279541), (u'caf\xe9_lattes', 0.4286726415157318), (u'coffee',
 0.42738500237464905), (u'BridgePort', 0.4086008369922638), (u'tripel', 0.4064597487449646), (u'joe', 0.4062146544456482), (u'fr
eshly_brewed', 0.39679768681526184), (u'espresso', 0.3963472247123718), (u'mixology', 0.3957287669181824)]
```

```
print w2vSO3M.most_similar(positive=['java', 'Django'], negative=['python'])
```

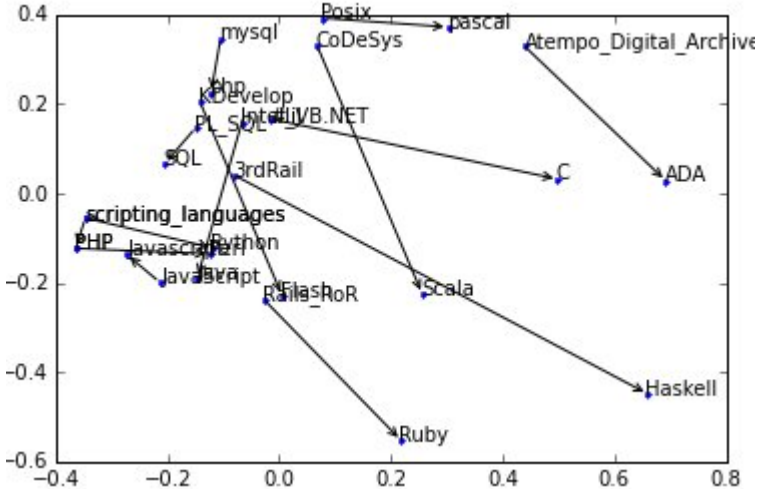
```
[(u'Grails', 0.5059091448783875), (u'django', 0.49170786142349243), (u'Wicket', 0.48384490609169006), (u'Spring', 0.477611780166
626), (u'GWT', 0.4758361577987671), (u'Vaadin', 0.4714881181716919), (u'CakePHP', 0.4712156355381012), (u'J2EE', 0.4463581442832
947), (u'Symfony', 0.44170787930488586), (u'Struts2', 0.4411844313144684)]
```

Uno propio vs uno general

IDE para cada Lenguaje
Stack Overflow



GoogleNews



¿En que he usado word2vec en mi trabajo?

Representación de la parte textual de los post.

El modelo word2vec obtiene una representación de cada palabra en un vector de tamaño 300.

Para modelar un post de SO calculó el promedio de los vectores de todas las palabras en el post.

Experimento:
Identificar relaciones entre los lenguajes de programación a partir de como las personas hablan de ellos en SO

Resultados de mi trabajo

	.net	acti	awk	bas	c	c++	c#	cloj	coff	colc	cud	curl	delp	f#	flex	go	groc	has	io	java	java	lua	mat	moc	obje	perl	php	pow	pyth	r	rub	sca	sed	swi	vba	vbscript
.net	33	0	0	1	1	3	6	1	0	7	2	0	4	2	3	1	0	1	1	1	6	0	1	5	5	0	1	3	3	0	0	3	1	0	3	2
actionscript	1	49	0	3	1	0	0	0	0	1	1	2	2	0	12	1	1	0	1	0	3	6	4	2	0	0	2	1	2	0	0	0	2	3	0	0
awk	0	0	52	10	0	1	0	2	2	0	0	0	0	2	0	1	0	0	0	0	0	0	1	0	0	0	1	0	2	5	0	0	19	0	1	1
bash	0	0	8	59	2	0	0	0	0	1	0	2	0	0	0	0	0	2	1	1	1	1	1	0	1	0	4	1	0	1	0	0	12	0	0	2
c	1	1	1	1	27	13	2	1	0	1	7	1	4	5	0	5	1	1	5	1	1	4	2	1	2	3	0	1	2	2	1	0	0	1	1	1
c++	6	0	0	1	16	28	1	0	0	4	3	1	2	0	1	3	0	5	8	1	0	4	1	1	2	1	0	2	1	1	2	0	0	2	1	2
c#	15	0	1	0	3	11	14	1	2	1	4	2	4	4	1	0	1	0	4	3	3	1	1	4	0	0	0	0	1	0	1	4	0	2	8	4
clojure	1	0	4	1	0	1	0	59	2	0	2	1	0	1	0	0	2	5	0	3	2	2	0	0	0	0	1	0	0	0	2	4	0	4	2	1
coffeescript	0	3	0	2	0	1	0	2	49	1	0	0	1	1	1	0	2	2	1	1	12	2	0	5	1	1	1	0	0	1	4	2	1	3	0	0
coldfusion	1	3	3	1	0	0	0	1	0	40	0	10	4	3	5	0	4	1	3	0	6	1	0	2	0	0	3	0	2	0	1	0	1	0	4	1
cuda	0	0	0	0	0	0	0	0	1	0	87	0	0	0	0	0	0	0	2	0	0	0	3	1	0	0	0	1	0	2	0	0	0	2	0	1
curl	0	1	0	0	0	0	0	0	1	1	0	85	0	0	0	0	1	0	0	0	2	0	0	0	2	0	5	0	1	0	0	0	0	0	0	1
delphi	0	1	1	1	2	0	0	0	1	4	2	1	55	1	1	1	1	0	10	1	0	1	0	0	6	0	0	3	0	0	1	0	0	3	1	2
f#	6	0	0	0	1	0	0	2	0	3	2	1	1	56	1	1	1	9	2	0	1	2	1	1	0	2	0	0	1	1	0	1	0	0	1	3
flex	0	13	1	1	1	1	0	0	0	1	0	1	3	0	56	1	1	0	3	3	4	0	1	0	1	0	2	0	1	0	0	0	0	2	3	0
go	0	2	2	2	1	0	0	4	1	1	1	3	0	1	1	64	0	1	2	0	1	1	0	4	0	1	0	2	0	0	2	0	2	1	0	0
groovy	0	0	1	1	1	0	0	1	2	3	0	4	0	1	2	2	50	3	1	8	1	2	0	4	2	0	0	1	0	1	3	3	0	0	3	0
haskell	0	0	0	1	2	0	0	2	3	0	1	1	2	6	0	3	0	54	4	1	1	3	1	0	1	1	0	1	0	3	2	1	0	5	1	0
io	1	2	0	3	1	3	0	0	1	0	3	1	0	0	0	3	0	9	47	3	2	0	2	1	4	4	1	1	2	0	1	0	3	1	0	1
java	2	2	0	2	1	4	0	2	2	3	3	1	3	3	0	3	4	0	9	37	1	0	0	3	2	0	3	0	1	1	2	3	0	1	2	0
javascript	1	1	0	0	0	0	0	0	3	0	0	1	0	0	1	0	1	0	0	3	73	0	0	3	1	1	0	1	4	1	2	1	1	0	0	1
lua	1	2	0	2	2	1	0	1	6	1	0	2	0	0	0	3	0	1	2	0	0	53	3	1	3	1	3	2	3	2	2	0	1	2	0	0
matlab	0	0	1	3	0	0	0	1	2	1	0	1	1	1	0	0	0	1	3	1	1	1	60	0	4	4	0	1	1	6	1	0	0	2	1	2
model	1	0	0	0	0	0	0	0	3	2	0	0	0	0	1	0	3	2	0	2	2	0	0	69	2	0	2	1	0	0	7	1	0	0	0	2
objective-c	0	4	1	0	0	3	1	0	2	0	2	2	2	1	1	1	0	1	1	1	0	0	0	0	66	0	0	1	0	0	1	0	0	1	0	0
perl	0	0	4	5	0	0	0	0	0	2	1	1	1	0	1	1	0	3	3	1	1	0	1	0	1	62	2	1	1	1	1	0	0	1	0	0
php	1	1	0	2	1	0	0	0	1	5	1	10	1	1	0	0	3	1	3	2	2	0	1	3	0	2	49	1	1	1	1	0	0	1	0	0
powershell	1	0	3	4	0	0	1	0	0	2	0	0	0	0	1	0	0	1	1	1	0	0	0	0	0	4	0	67	1	2	1	0	0	1	0	0
python	0	0	2	5	1	1	1	0	3	1	0	0	1	0	1	2	1	4	3	2	1	1	2	3	1	4	3	1	47	3	5	0	0	0	1	0
r	1	3	2	0	0	1	0	1	0	0	1	0	0	1	0	1	0	1	1	1	1	0	7	0	0	0	0	1	1	74	0	0	1	0	1	0

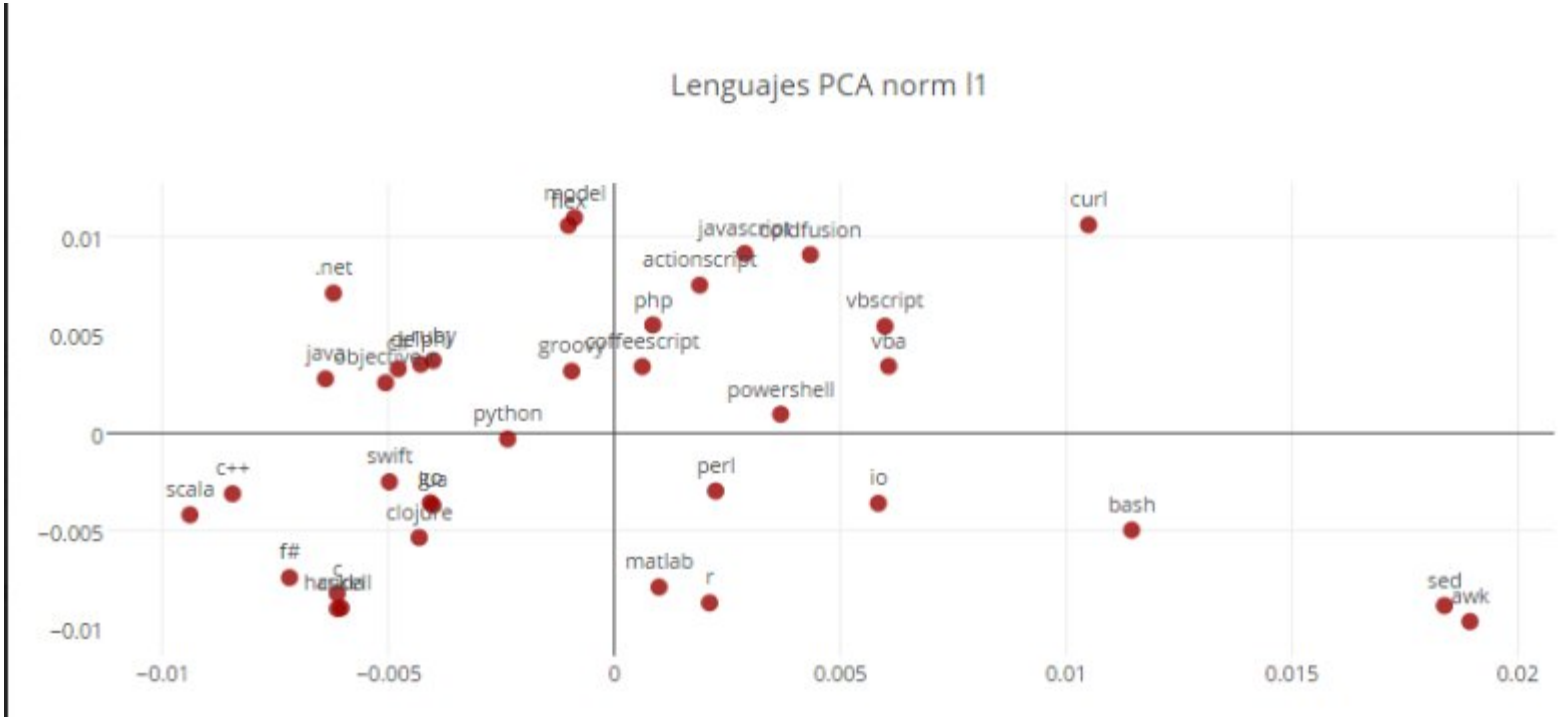
```
f1 0.530046329668
acc 0.537520844914
Prec 0.53005598986
recal 0.537520844914
```

Resultados de mi trabaio

	java	python	php	c#	javascript	c++	c	objective-c	r	swift	matlab	ruby	vb.net	vba	perl	scala	lua	delphi
java	60	2	5	1	5	3	1	2	0	2	2	0	1	1	2	7	2	4
python	2	51	2	0	4	2	5	2	3	4	1	6	1	2	7	4	3	1
php	1	2	75	2	5	1	0	1	2	0	2	1	0	1	4	0	1	2
c#	2	0	2	32	6	8	7	6	1	3	3	1	11	5	4	3	1	5
javascript	1	5	5	1	76	0	0	3	0	0	0	1	2	0	2	1	0	3
c++	4	2	2	7	1	25	31	2	1	2	3	2	4	0	3	2	5	4
c	3	3	0	4	4	10	39	0	3	4	6	1	2	0	8	2	1	10
objective-c	1	4	3	1	2	3	2	55	0	18	3	1	0	2	2	1	1	1
r	2	2	3	0	0	0	0	1	74	0	9	3	0	1	2	1	1	1
swift	0	1	1	2	2	5	3	17	1	56	2	1	0	2	1	4	0	2
matlab	2	3	1	1	2	2	2	1	9	2	68	0	0	3	0	1	1	2
ruby	1	3	4	2	5	0	0	3	3	0	1	72	0	1	2	2	1	0
vb.net	2	1	1	9	2	3	0	2	1	1	0	2	61	7	1	1	1	5
vba	1	0	0	0	1	2	0	0	2	1	1	1	2	82	1	1	0	5
perl	1	3	7	1	2	1	1	0	4	3	1	4	1	4	63	0	4	0
		3	1	2	2	1	3	0	3	5	1	3	0	0	3	64	1	0
		4	2	1	1	4	5	0	3	4	3	2	1	2	5	1	60	1
		2	3	4	0	2	1	0	2	1	5	1	2	0	1	0	1	74

```
f1 0.597544326066
acc 0.603888888889
Prec 0.597324061389
recal 0.603888888889
```

Resultados de mi trabajo



Extra

¿Como entrenar modelos word2vec?

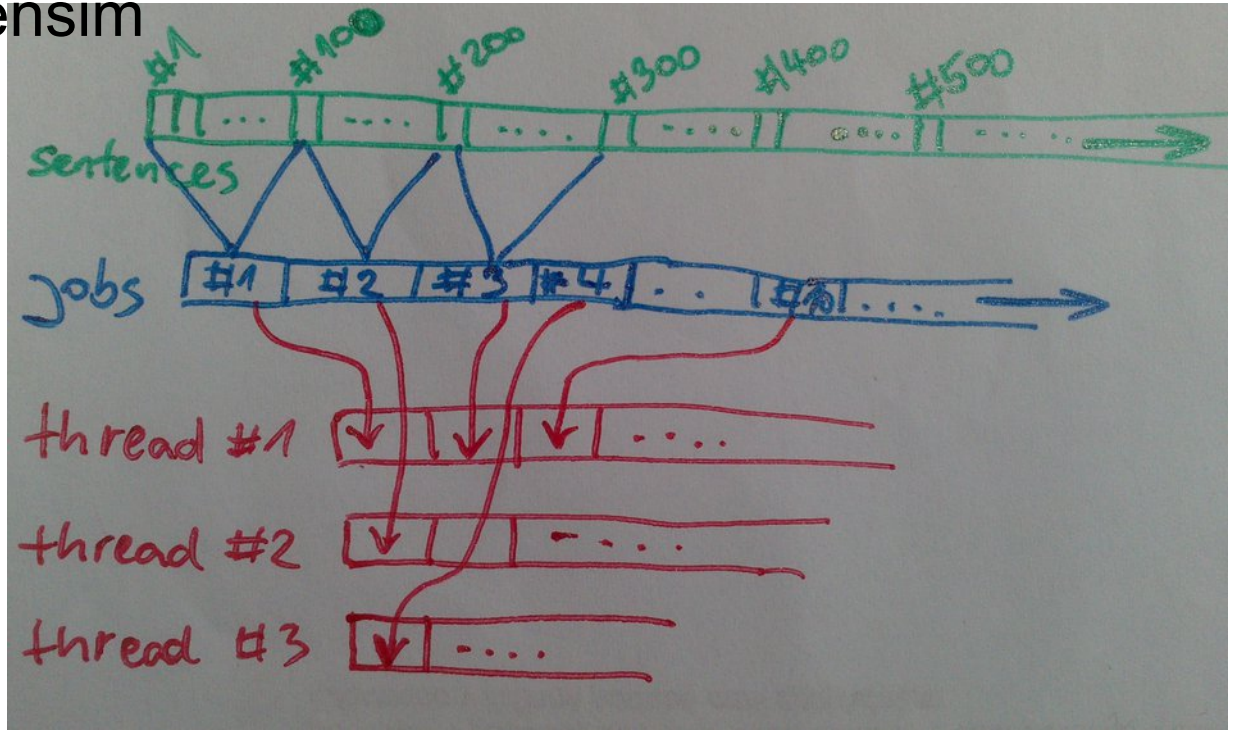
Usando Gensim

```
import gensim
from gensim.models import word2vec
#Documentation: https://radimrehurek.com/gensim/models/word2vec.html

num_features = 300    # Word vector dimensionality
min_word_count = 5    # Minimum word count
num_workers = 16      # Number of threads to run in parallel
context = 5           # Context window size
downsampling = 0      # Downsample setting for frequent words
```

Extra

¿Como entrenar modelos word2vec?

Usando Gensim



Extra

¿Como entrenar modelos word2vec?

Usando Gensim

Al Modelo Word2Vec se le pasan los parametros y una lista o Iterador con las Oraciones separadas por tokens las cuales se usan para entrenar el modelo.

```
print "Training model..."
model = word2vec.Word2Vec(sentences, workers=num_workers, \
        size=num_features, min_count = min_word_count, \
        window = context, sample = downsampling)
print "Fin training model."
```

```
Training model...
Fin training model.
```

Extra

¿Como entrenar modelos word2vec?

Usando Gensim

¿Cómo separar las Oraciones? ¿Como separar los Tokens?

```
[u'I', u'want', u'to', u'use', u'a', u'track', u'-', u'bar', u'to', u'change', u'a', u'form', u"'", u's', u'opacity', u'.']
[u'This', u'is', u'my', u'code', u':']
[u'When', u'I', u'try', u'to', u'build', u'it', u',', u'I', u'get', u'this', u'error', u':']
[u'Cannot', u'implicitly', u'convert', u'type', u"'", u'decimal', u"'", u'to', u"'", u'double', u"'."]
[u'I', u'tried', u'making', u'a', u',', u'but', u'then', u'the', u'control', u'doesn', u"'", u't', u'work', u'.']
[u'This', u'code', u'has', u'worked', u'fine', u'for', u'me', u'in', u'VB', u'.', u'NET', u'in', u'the', u'past', u'.']
[u'I', u'have', u'an', u'absolutely', u'positioned', u'containing', u'several', u'children', u',', u'one', u'of', u'which', u'i
s', u'a', u'relatively', u'positioned', u'.']
[u'When', u'I', u'use', u'a', u'percentage', u'-', u'based', u'width', u'on', u'the', u'child', u',', u'it', u'collapses', u't
o', u'width', u'on', u'IE7', u',', u'but', u'not', u'on', u'Firefox', u'or', u'Safari', u'.']
[u'If', u'I', u'use', u'pixel', u'width', u',', u'it', u'works', u'.']
[u'If', u'the', u'parent', u'is', u'relatively', u'positioned', u',', u'the', u'percentage', u'width', u'on', u'the', u'child',
 u'works', u'.']
[u'Is', u'there', u'something', u'I', u"'", u'm', u'missing', u'here', u'?']
[u'Is', u'there', u'an', u'easy']
[u'Are', u'there', u'any', u'conversion', u'tools', u'for', u'porting', u'from', u'Visual', u'J', u'#', u'code', u'to', u'C',
 u'#', u'?']
```

Extra

¿Como entrenar modelos word2vec?

Usando Gensim
¿Cómo separar las Oraciones?

La forma facil usando nltk:
Los corpus de nltk se puede extraer las sentencias.
Para corpus no incorporados en nltk se puede usar PlaintextCorpusReader

```
from nltk.corpus import PlaintextCorpusReader
sentences = nltk.corpus.gutenberg.sents(u'bible-kjv.txt')
wordList = PlaintextCorpusReader('', 'sentences1.txt')
```

Extra

¿Como entrenar modelos word2vec?

Usando Gensim
¿Como separar los Tokens?

¿nltk al rescate?
nltk.word_tokenize(texto)

[u'Are', u'there', u'any', u'conversion', u'tools', u'for', u'porting', u'from', u'Visual', u'J', u'#', u'code', u'to', u'C', u'#', u'?']

¿Es lo que queremos?

No! En casos como este hay que realizar una tokenización más cuidadosa.

PlaintextCorpusReader retorna las sentencias dividas en tokens.

Extra

¿Como entrenar modelos word2vec?

Usando Gensim
¿Si el corpus no cabe en memoria?
Usar un iterador (Tener cuidado)

```
class MySentencesHDF5(object):
    def __init__(self, hdf5TableInfo):
        self.table = hdf5TableInfo

    def __iter__(self):
        start = time()
        for row in self.table:
            sentences = toSentencesArray(row[self.atrib], self.atrib)
            if sentences:
                for sent in sentences:
                    yield sent
```

¿Si quiero continuar entrenandolo?

¡Gracias!

Preguntas en: