



Text2VecClust unsupervised data exploration

Task: EXATEL - Data categorizing software

BION

Maciej Sykulski

Weronika Kolczyńska

Jakub Grabowski



Software used

- Python - library **Gensim**:
 - Scalable statistical semantics
 - Analyze plain-text documents for semantic structure
 - Retrieve semantically similar documents
 - We use it to submerge word space into 150 dimensional vector space
- GIT
- R
 - we approached the word2vec analysis in R with text2vec package, however it produced less intuitive word2vec associations than Python gensim package, tokenization method could be one factor, but we suspect that not only
- Jupyter
 - used for analysis and data exploration

Word2Vec trained on 1/15th of the whole data - similarities

```
In [8]: w1 = ["beyonce"]  
model.wv.most_similar_cosmul(positive=w1,topn=10)
```

```
Out[8]: [('timberlake', 0.8396907448768616),  
( 'minaj', 0.8058559894561768),  
( 'justin', 0.7937102913856506),  
( 'album', 0.7908747792243958),  
( 'cardi', 0.7862465977668762),  
( 'billie', 0.7852219939231873),  
( 'nicki', 0.781724750995636),  
( 'feat', 0.7798829078674316),  
( 'meghan', 0.7768550515174866),  
( 'jay', 0.7746703624725342)]
```

```
In [17]: w1 = ["france"]  
model.wv.most_similar_cosmul(positive=w1,topn=10)
```

```
Out[17]: [('germany', 0.8786470293998718),  
( 'europe', 0.8778382539749146),  
( 'spain', 0.8593621850013733),  
( 'paris', 0.8377141356468201),  
( 'canada', 0.8344048261642456),  
( 'russia', 0.8334130644798279),  
( 'denmark', 0.8285591006278992),  
( 'greece', 0.8272826075553894),  
( 'obstétriciens', 0.8263376121521),  
( 'switzerland', 0.8231014013290405)]
```

```
In [10]: w1 = ["pokemon"]  
model.wv.most_similar_cosmul(positive=w1,topn=10)
```

```
Out[10]: [('pokémon', 0.8885942101478577),  
( 'pikachu', 0.7988919615745544),  
( 'salamence', 0.7925139665603638),  
( 'darkrai', 0.7895944714546204),  
( 'alakazam', 0.7876683473587036),  
( 'groudon', 0.7824632525444031),  
( 'psychic', 0.782210111618042),  
( 'lunala', 0.7822090983390808),  
( 'abrave laugh', 0.781223535377197),  
( 'ditto', 0.7809909582138062)]
```

```
In [26]: w1 = ["computer"]  
model.wv.most_similar_cosmul(positive=w1,topn=10)
```

```
Out[26]: [('computers', 0.8517665266990662),  
( 'laptop', 0.8027507662773132),  
( 'systems', 0.7939674258232117),  
( 'smartphones', 0.7836009860038757),  
( 'machine', 0.7801939249038696),  
( 'software', 0.7796436548233032),  
( 'phones', 0.7786757349967957),  
( 'devices', 0.7716999053955078),  
( 'hardware', 0.7674819827079773),  
( 'device', 0.7650493383407593)]
```

```
In [18]: w1 = ["king"]  
model.wv.most_similar_cosmul(positive=w1,topn=10)
```

```
Out[18]: [('prince', 0.8689006567001343),  
( 'lord', 0.8523561358451843),  
( 'knight', 0.8440752029418945),  
( 'queen', 0.8399626612663269),  
( 'prophet', 0.8229497075080872),  
( 'dynasty', 0.8226714134216309),  
( 'emperor', 0.8204849362373352),  
( 'satan', 0.8175635933876038),  
( 'kings', 0.8168355822563171),  
( 'captain', 0.8132425546646118)]
```

```
In [11]: w1 = ["food"]  
model.wv.most_similar_cosmul(positive=w1,topn=10)
```

```
Out[11]: [('eating', 0.8694721460342407),  
( 'drinks', 0.8657129406929016),  
( 'drink', 0.8479171395301819),  
( 'foods', 0.845889687538147),  
( 'eat', 0.8446840047836304),  
( 'meals', 0.8358771800994873),  
( 'cooking', 0.8339307904243469),  
( 'alcohol', 0.831096351146698),  
( 'snacks', 0.8296454548835754),  
( 'fish', 0.8294864296913147)]
```

```
In [237]: model.wv.most_similar_cosmul(positive=['woman', 'king'], negative=['man'], topn=1)
```

```
Out[237]: [('queen', 0.8711902499198914)]
```

inference after adding or subtracting words

```
In [221]: model.wv.similarity(w1="men", w2="woman")
```

```
Out[221]: 0.5107145
```

```
In [228]: model.wv.similarity(w1="chicken", w2="nuggets")
```

```
Out[228]: 0.45717996
```

```
In [229]: model.wv.similarity(w1="hau", w2="miau")
```

```
Out[229]: 0.55619955
```

```
In [231]: model.wv.similarity(w1="small", w2="little")
```

```
Out[231]: 0.4925781
```

```
In [214]: model.wv.similarity(w1="dirty", w2="clean")
```

```
Out[214]: 0.22903207
```

```
In [219]: model.wv.similarity(w1="dirty", w2="filthy")
```

```
Out[219]: 0.45736545
```

```
In [220]: model.wv.similarity(w1="dirty", w2="dirty")
```

```
Out[220]: 1.0
```

(These similarities are computed with different formula than the ones on the previous page)



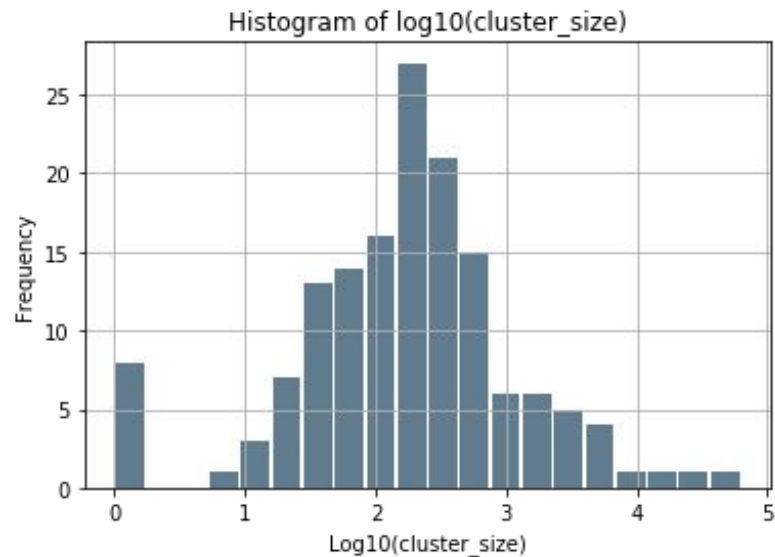
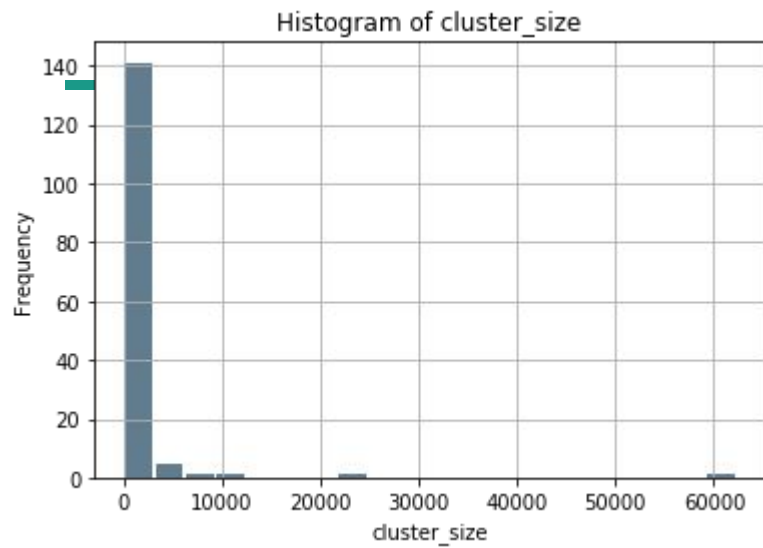
Text2VecClust

Data categorizing tool which is based on clustering according to words occurring in text

General steps:

- Parsing files into words and creating a list of them
- Creating and training model based on the list
- Submerge text fragments into vector space in the following way.
- We concatenate:
 - average of all word vectors from text fragment
 - difference between sum(average didn't work well) of word vectors from the 1st 1/4th of text fragment and last 3/4th
 - difference between sum of word vectors from the 1/2nd of text fragment and last 1/2nd
 - difference between sum of word vectors from the 3/4th of text fragment and last 1/4th
- We clusterize these six hundred dimensional using K-means algorithm
- In our plans we had SVD decomposition of this high-dimensional matrix with irlba package in R

Clustering results



Cluster examples - the largest cluster

Source code but not only

```
[[["$menu_name = 'menu'\n",
'$locations = get_nav_menu_locations();\n',
'$menu = wp_get_nav_menu_object( $locations[ $menu_name ] );\n',
"$menuitems = wp_get_nav_menu_items( $menu->term_id, array( 'order' => 'DESC' ) );\n",
'\n',
'foreach( $menuitems as $item):\n',
'\t\t$title = $item->title;\n',
'\t\t$link = $item->url;\n',
'endforeach;'],

["Cscript %windir%\system32\SCRegEdit.wsf /ar 0" and press the "Enter" button.],

['local GUI = require("GUI")\n',
'local MineOSInterface = require("MineOSInterface")\n',
'local image = require("image")\n',
'\n',
'\n',
'local mainContainer, window = MineOSInterface.addWindow(GUI.titledWindow(50, 30, 70,
27, "Hello,World 2!", true))\n',
'window.backgroundPanel.colors.transparency = 0.2\n',
'\n',
'\n',
'window.addChild(GUI.label(17, 10, window.width, window.height, 0x5A5A5A, "Hello,World!
2 (Windowed Edition(Fixed))")\n',
'window.addChild(GUI.label(2, 26, window.width, window.height, 0x5A5A5A, "Version
1.04"))\n',
'window.addChild(GUI.label(48, 26, window.width, window.height, 0x5A5A5A, "by
Sergo73,2018 MIT"))'],

['# Query_time: 20.829720 Lock_time: 0.000169 Rows_sent: 7282550 Rows_examined:
14604241\n',
'# Rows_affected: 0\n',
'use affray;\n']
```

```
['1. Instalacja Dockera\n',


'https://docs.docker.com/engine/installation/linux/docker-ce/ubuntu/#install-docker-ce\n',
'2. Instalacja Docker Compose\n',
'https://docs.docker.com/compose/install/\n',

'https://docs.docker.com/engine/installation/linux/docker-ce/ubuntu/#install-docker-ce\n',
'\n',
'3. Instalacja środowiska\n',
'https://gorails.com/setup/ubuntu/16.04\n',
'\n',
'*Uwaga: wybrać opcję rvm*\n',
'Do punktu *Installing Rails*, bez instalacji baz
danych\n',
'\n',
'\n',

'https://docs.docker.com/engine/installation/linux/linux-postinstall/#manage-docker-as-a-non-root-user\n',
'\n',
'sudo service docker start - raz, potem się
powinno włączać samo'],
[{'\n',
'\t"editor.colorDecorators": false,\n',
'\t"editor.detectIndentation": false,\n',
'\t"editor.formatOnSave": false,\n',
'\t"editor.fontSize": 14,\n',
'\t"editor.insertSpaces": false,\n',
'\t"editor.matchBrackets": false,\n',
'\t"editor.renderLineHighlight": "gutter",\n',
'\t"editor.wordWrap": "on",\n',
'\t"git.autofetch": true,\n',
'\t"git.confirmSync": false,\n',
```

Cluster examples

cluster of size 51



```
['Booop-bop bop! bip-bip bop! biiiiip boop! bip! bip booooop biiiiip\n',  
  'Bip booooop bop booooop-bop biiiiip boop!\n',  
  'Bip bop! bip-bip bop\n',  
  'Bip! bip! bip biip bip-bip bip bop bip biip bip-bip bip bip booooop boop! bop bip boop! boop! biip booooop bip! bip boop!\n',  
  'Bop bip! bop! bop! booop-bop bip-bip bip booooop bip-bip bop biip biiiiip bip! bop booooop bop bop biiiiip biiiiip bip! bip booop-bop bop biiiiip bop! bop!  
booooop\n',  
  'Bip bop biip bop! bip? bop boop! booooop bop bop bip! bop? boop! booooop biiiiip bop booooop bop! boop! biiiiip biip biip bop! booooop bop? bop? bop bip?  
biiiiip bop! bop? bop? booooop\n',  
  'Bop booop-bop bop? bop bop bip-bip boop! biiiiip bip-bip bip bop booop-bop bop? bip-bip boop! biiiiip bip-bip bip! bip-bip bop bip! bop? bip! biip bip?\n',  
  'Boop! bip! bop bop bop bip-bip  
...]
```

```
['Bop? bop? bip biip bop bip-bip\n',  
  'Booooop booop-bop bip? bip! bip bop boop! bip-bip bop bip-bip bop? boop! bip! bip! biip bop bop? bop bop? booooop bip bip-bip biip biip  
bop? bop! bop bip booooop booooop biip biiiiip bop biip bip\n',  
  'Booop-bop bip biip bip? bip biip bip! bop bop! bip bip-bip booooop bip? booooop booop-bop booop-bop\n',  
  'Bop bip bop? bop bop! booop-bop bop biiiiip bip! bop bip! bop! biip bop bip? bip-bip bop! boop! boop! bip? biip bip bip bop biip bip bip  
bop bip? boop! bop? bip-bip bip! bip! bip! biip bip-bip booooop biiiiip bop bip\n',  
  'Bip bop? bip biip bip bip-bip booooop bip bip-bip bop bip! booooop biip biip bop biip booooop bop biiiiip bip! bip? bop? biip bip-bip bip bip  
bop bip! bop bip bop bip-bip bip-bip bip? bop booooop bop? bip bop bip booooop bop bip biip bop? bip! bip?\n',  
  'Booooop bop booop-bop bip! bop! biiiiip bop! boop! bip! boop! bip booooop biiiiip bip bop! bip bop bip biip bip! biip bip? booop-bop  
bop! booooop boop! bop? bop! bop\n',  
  'Bip booooop bip? bop booop-bop boop! bop boop! bop bip! bop? bip-bip bip! booop-bop boop! bop? bop! biiiiip bop? booop-bop bop!  
biiiiip booop-bop booooop bop? bip-bip bip bop? bip? bip! bop bip-bip bop?\n',  
  'Bip biiiiip bip booooop biiiiip bip-bip bop bop? biiiiip bip! bop bip?\n',  
...]
```


Cluster examples

[['Daemons burned in their thousands, their aetheric flesh seared from their false bones. White flame haloed from the sword in corrosive, purifying radiance. It coruscated in thrashing waves from each fall of the Emperor's blade. To look at Him was to go blind. To stand before Him was to die.]\n',

'http://wh40k.lexicanum.com/wiki/The_Master_of_Mankind_(Novel)\n',

\n',

cosmic/game/demon
texts

['The Imperial Creed and Spirits of the Immaterium\n',

'The official position of the Ecclesiarchy on the spirits of the deceased is that the Emperor judges all faithful humans after death and, if they are worthy, grants them a place in his celestial army. Differing interpretations of the Imperial Creed offer a wide variety of explanations for what happens to those souls deemed unworthy of joining the God-Emperor's ranks, but who are not so heretical as to be damned out of hand. Some versions say they are reborn to try again, others, that they must wander the afterlife for a time, braving the dangers of the warp as penance for a life ill spent until their actions have redeemed them, proving them worthy of the God-Emperor's service. There are also many tales of legendary servants of the Emperor returning from the immaterium to the world of the living when the people of the Imperium once again need them. Some versions of the Creed refuse to acknowledge the sentience of such entities, referring to them in technical terms such as "post-life warp signatures" and "the aetheric charge contained by a residual personality". Regardless of the fine points of doctrine, the Ecclesiarchy does acknowledge the existence of spirits of the dead. Several branches of the Inquisition take a very active interest in such entities and their relationship to the warp. The bulk of Ordo Xenos,...]

\n',

"[The afterlife was a perilous place. Souls teemed and whirled in great shoals upon its currents. Some might return to corporeal existences, others became things greater or lesser than the beings they had been. Many more were torn to shreds by the warp's voracious predators. Others simply faded to nothing.]\n"],

['\n',

"Our scout checked the direct route; she said it looked clear at first, but noticed something big shadowing her from under the trees at sunset. This is a bad omen. We should find a way around here."\n',

\n',

'History\n',

\n',

'These gryphons were always outsiders even to the other gryphons, usually due to their seclusive behavior, but in current day, it is also because of their aggressiveness and in some cases, outright predatory behavior that has found dragons and even other gryphons alike the subject of being hunted.\n',

\n'],

['\n',

'1.1 Faction leaders \n',

\n',

'Sith Emperor\n',

\n',


'As the sith emperor you have full control of the day to day running and activities of the sith order, You may select your high ranking officials and great a higharach \n',

'in order to maintain your rain. You are fully responsible for the actions of members of the sith order and are expected to rule with an iron fist. \n',

'Upon successful overthrow and claiming the title of emperor, You may promote any member of your choosing for the position of Emperors Voice, Emperors Hands. Please note Emperors\n',

Cluster examples

stories



```
['I love waking up to drama.\n',  
 '\n',  
 'ValeneToday at 12:14 PM\n',  
 'Jay\n',  
 'I'm going to tell you right now\n',  
 '\n',  
 'Jay™Today at 12:15 PM\n',  
 'Rachel asked me to not speak to you again.\n',  
 '\n'], [...]
```

```
["setting. Reporting in from base, from the uh command post, i.e. bed, but I-uh, oh man it's\n",  
 "been one of those days. Y'know, it's been one of those days, cause I said originally I was\n",  
 "going to do a review today but it's not gonna happen, I'm gonna tell you. I think it's a\n",  
 "good chance to kinda set the camera up, y'know do it maybe a little bit casually and uh\n",  
 'just try to, just try to vent it out to you and maybe get a message across at the same\n',  
 "time. Ugh, I mean don't even get me started, alright, it was one of those days. You know\n",  
 'when just everything goes wrong and you have absolutely no control over the situation\n',  
 "whatsoever, one of those days where y'know I'm filming this on a Thursday right, so it's\n",  
 "not live but it's ok. Y'know it's thursday evening it's around 7:18PM. For those of you\n"], [...]
```

```
['\n',  
 '\tYou roll right, pulling back on the stick. The light Arwing lines itself up with the fighter in front and you pull the trigger. Two green beams come out of your  
laser cannons, hitting their mark and destroying the fighter in front of you. You hear a loud beeping and proceed to pull back on the stick, breaking the lock.\n',  
 '\n',  
 [...]
```

Cluster examples

encoded data

```
[[['AEAAHAAbAAYAAAAEAAAAQAAAAAAAAABAAAAAAAAAAEAAAAAAAAAAaAIAAAAAABoAgAAAAAAAAGa\n',  
  'AAAAAAAAAwAAAAQAAACoAgAAAAAAAKgCAAAAAAAAgAIAAAAAAAAcAAAAAAAABwAAAAAAAQA\n',  
  'AAAAAAAABAAAABAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAJg+AAAAAAAmD4AAAAAAAEEAAA\n',  
  'AAAAAEAAAAFAAAAAEAAAAAAAAAAQAAAAAAAAABAAAAAAAAAaeUBAAAAABp5QEAAAAAAAQAAAA\n',  
  'AAAAAQAAAAQAAAAAMAIAAAAAAAAAwAgAAAAAADACAAAAADYpwAAAAAAANinAAAAAAAAABAAAA\n',  
  'AABAAAAABgAAADDiAgAAAAAAMPICAAAAAAAw8glIAAAAAAPgXAAAAAAAACDcAAAAAAAEEAAAA\n',  
  'AAIAAAGAAAAACoKAAAAAAAI+QIAAAAAAJ5AgAAAAA8AEAAAAAADwAQAAAAAAGAAAAAA\n',  
  'BAAAAAQAAADEAgAAAAAAMQCAAAAAAAAxAlIAAAAAABEAAAAAAAAAEQAAAAAAAABAAAAAAAABQ\n',  
  '5XRkBAAAAHR/AgAAAAAdH8CAAAAAAB0fwlIAAAAAOQMAAAAAAA5AwAAAAAAAEEAAAAAAAFH\n',  
  ...],  
 [['AEAAHgAdAAYAAAAFAAAAQAAAAAAAAABAAAAAAAAAAEAAAAAAAAAA+AEAAAAAAD4AQAAAAAAAAGa\n',  
  'AAAAAAAAAwAAAAQAAAA4AgAAAAAADgCAAAAAAAOAIAAAAAAAcAAAAAAAABwAAAAAAAQA\n',  
  'AAAAAABAAAAABQAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAABx6AgAAAAAAHhOCAAAAAAAACAA\n',  
  'AAAAAEAAAAAGAAAASHwCAAAAAABIfCIAAAAAAEh8IgAAAAAAuRoAAAAAABBIwAAAAAAAAlAAA\n',  
  'AAAAAgAAAAYAADgfQIAAAAAAOB9IgAAAAA4H0iAAAAAADwAQAAAAAAPABAAAAAAACAAAAAA\n',  
  'AAEAAAAABAAAAFQCAAAAAAAVAIAAAAAABUJgAAAAAAEQAAAAAAAAARAAAAAAAAAEAAAAAA\n',  
  'AFDIdGQEAAAA1DUCAAAAAADUNQIAAAAAANQ1AgAAAAAAhAkAAAAAAACECQAAAAAAAQAAAAAA\n',  
  'UeV0ZAYAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAEAAAAAAAABS\n',  
  '5XRkBAAAAEh8AgAAAAAASHwiAAAAAABIfCIAAAAAALgDAAAAAAAUAMAAAAAAAABAAAAAAAC9s\n',  
  ...],  
 [['AEAGwAaAAYAAAAEAAAAQAAAAAAAAABAAAAAAAAAAEAAAAAAAAAAaAIAAAAAABoAgAAAAAAAAGa\n',  
  'AAAAAAAAAwAAAAQAAACoAgAAAAAAKgCAAAAAAAAgAIAAAAAAAAcAAAAAAAABwAAAAAAAQA\n',  
  'AAAAAABAAAABAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAChLAAAAAAAKEsAAAAAAAEEAAA\n',  
  'AAAAAEAAAAFAAAAAFAAAAAUAAAAAAAAABQAAAAAAAAAPWEBAAAAAA9YQEAAAAAAAQAAAA\n',  
  'AAAAAQAAAAQAAAAAwAEAAAAAADAAQAAAAAAMABAAAAAADIZgAAAAAAMhmAAAAAAAABAAAA\n',  
  'AABAAAAABgAAAMg1AgAAAAAYEUCAAAAAADIRQIAAAAAAGPAAAAAAAOBAAAAAAAEEAAAA\n',  
  'AAIAAAGAAAA2DUCAAAAAADYRQIAAAAAANhFagAAAAAAYAlIAAAAAABgAgAAAAAAGAAAAAA\n',  
  'BAAAAAQAAADEAgAAAAAAMQCAAAAAAAAxAlIAAAAAABEAAAAAAAAAEQAAAAAAAABAAAAAAAABQ\n',  
  '5XRkBAAAAOj7AQAAAAAA6PsBAAAAADo+wEAAAAAOwFAAAAAAA7AUAAAAAAAEEAAAAAAAFH\n',  
  ...]]
```

Cluster examples

logs

```
[[[11:05:01.879] [Info] Root: Done preparing Root.\n',  
'[11:05:01.879] [Info] Client Version 1.3.4 (windows x86_64) Source ID: f468c99f113f3d44bb548367da9c9a1c6f4b031d Protocol: 745\n',  
"[11:05:01.879] [Info] Root: Scanning for asset sources in directory '..\\assets\\'\n",  
"[11:05:02.370] [Info] Root: Scanning for asset sources in directory '..\\mods\\'\n"],  
[...]  
['[13:21:44.939] [Info] Root: Done preparing Root.\n',  
'[13:21:44.939] [Info] Client Version 1.3.3 (windows x86_64) Source ID: c21069b204d626bfe673a047a2100d2fcd0766d4 Protocol: 743\n',  
"[13:21:44.939] [Info] Root: Scanning for asset sources in directory '..\\assets\\'\n",  
"[13:21:45.224] [Info] Root: Scanning for asset sources in directory '..\\mods\\'\n"],  
[...]  
['[20:47:11.451] [Info] Root: Done preparing Root.\n',  
'[20:47:11.451] [Info] Client Version 1.3.3 (macos x86_64) Source ID: c21069b204d626bfe673a047a2100d2fcd0766d4 Protocol: 743\n',  
"[20:47:11.451] [Info] Root: Scanning for asset sources in directory '..\\assets\\'\n",  
"[20:47:11.867] [Info] Root: Scanning for asset sources in directory '..\\mods\\'\n"],  
[...]  
['[22:02:23.013] [Info] Root: Done preparing Root.\n',  
'[22:02:23.013] [Info] Client Version 1.3.4 (windows i386) Source ID: f468c99f113f3d44bb548367da9c9a1c6f4b031d Protocol: 745\n',  
"[22:02:23.015] [Info] Root: Scanning for asset sources in directory '..\\assets\\'\n",  
"[22:02:23.826] [Info] Root: Scanning for asset sources in directory '..\\mods\\'\n"],
```

Cluster examples

thematic posts about hacking

```
[['\n',  
    Latest Working Grand Theft Auto 6 Hack\n',  
    '\t\t\t\t\t \n',  
    '\t\t\t\t\t Secret Link:\n',  
    '\t\t\t\t\t http://bit.ly/2wq8GH3\n',  
    '\n',  
    '\n',  
    '\n',  
    [...],  
    '\n',  
    Latest Working NBA Hack\n',  
    '\t\t\t\t\t \n',  
    '\t\t\t\t\t Secret Link:\n',  
    '\t\t\t\t\t http://bit.ly/2PjDINo\n',  
    '\n',  
    '\n',  
    '\n',  
    [...],  
    '\n',  
    Latest Working Musically Hack\n',  
    '\t\t\t\t\t \n',  
    '\t\t\t\t\t Secret Link:\n',  
    '\t\t\t\t\t http://bit.ly//2wwQsD7\n',  
    '\n',  
    '\n',  
    '\n',  
    [...],  
    '\n',  
    Latest Working Fortnite Hack\n',  
    '\t\t\t\t\t \n',  
    '\t\t\t\t\t Secret Link:\n',  
    '\t\t\t\t\t http://bit.ly/2LF7g05\n',  
    '\n',  
    '\n',  
    '\n',  
    [...]]]
```

Cluster examples

```
'\n',
'#EXTINF:-1 tvg-id="Cancao Nova" tvg-logo="https://i.imgur.com/CSoMkJU.png?1" group-title="CANAIS DE TV",CAN❖❖O NOVA\n',
'http://tvajuhls-lh.akamaihd.net:80/i/tvdesk_1@147040/index_1080_av-p.m3u8\n',
'\n',
[...]  
" 'object'\n",  
" 'float64'\n"],  
['[18:36:45] <@^Botela_> terrortool Labas HeyGuys\n',  
'[18:42:42] <TerrorTool> kiek priesu dar like?\n',  
'[18:43:10] <TerrorTool> oba tik paklausiau ir parode ekrane\n',  
'[18:43:26] <TerrorTool> niekas nemire kol nebuva, bet mire va katik\n',  
'[18:43:53] <TerrorTool> rode 34 ir 15, jei spejau perskaityt tai 15mire\n',  
'[18:45:07] <TerrorTool> mire ir i lavoneliu istaiga isiusti buvo\n',  
'[20:53:56] <TerrorTool> cia norway bus pietine dalis\n',  
'[20:57:06] <TerrorTool> lmaoo... keli tukstanciai km :D gg well played sponsi\n',  
'[22:42:51] <TerrorTool> @Sponsorius yra.... pats pirmas\n',  
'[22:44:31] <TerrorTool> Tbili sisi\n',  
'[22:51:46] <TerrorTool> laba svakara schebra ir @Sponsorius turiu klausima, gal kuris zinosit\n',  
"[22:53:09] <TerrorTool> kai spaudi pas streameri donate ir atvercia paypal'a, kaip paypal ta mokejima mato, ar pay for service ar send to friends... ?\n",  
[...]  
['title: Dynamic Component Templates với Vue.js\n',  
'author: PhongPV\n',  
'date: 2018-06-29T23:22+07:00\n',  
"tags: ['code']\n",  
'--\n',  
'\n',  
'> Components không phải luôn luôn có cùng cấu trúc. Đôi lúc chúng có nhiều states khác nhau để quản lý. Nó sẽ thích hợp để dùng trong trường hợp bất đồng  
bộ.\n',  
'\n',  
'## Những trường hợp sử dụng\n',  
'\n',  
'- Component templates được sử dụng trong Scrumpy và trong hầu hết các vị trí như thông báo, nhận xét và tập đính kèm. Hãy nhìn một số _nhận xét_ và nhìn  
chúng thực sự nghĩa là gì.\n',  
'\n',  
[...]
```

foreign language chat and link lists

Discussion



Results of clustering are surprisingly partially good, clusters exhibit interesting similarities. It's difficult to assess if large clusters are homogenic enough. The idea of summing word vectors from text fragments is based on the target vector space being large enough to contain text fragment types. For larger texts this assumption may fail, and sums of many diverse words might produce less pronounced vector directions, thus resulting with less homogenic clustering. To verify if this is the case one would need to measure homogeneity of clusters.

Development

Search or filter results...

Edit board

Add list

Add issues

Open 0

To Do 11

Doing 0

1st category tasks 4

GOAL categorize unlabelled data with given input

1st category tasks To Do

#8

GOAL Allow some way to browse the resulting data by assigned category (filter category IN or OUT)

1st category tasks To Do

#9

GOAL determine parameters for evaluation, for e.g. "let's assume document is in category X if probability is higher than 66%"

1st category tasks To Do

#16

GOAL evaluate algorithm on 100% of data and cach the result

1st category tasks To Do

#10

2nd category task 3

GOAL see category data of a document given by ID

2nd category task To Do

#13

GOAL list of documents matching a given category and compare their contest to see how weel they fit the group (additional sort by how well they fit the category)

2nd category task To Do

#12

GOAL manual analysis, select the most common categories and "name" them (it will help to evaluate results))

2nd category task To Do

#14

3rd category additional 2

GOAL dedicated web interface

3rd category additional To Do

#11

THINK what is better - categorise whole documents or split them into smaller parts and categorise parts instead

3rd category additional To Do

#15

Closed 5

Go to conference at 2000

#6

read on word2vec

#7

experiment with word2vec

#3

Apply word2vec

#1

Read part of the data into python

#2



**Thanks
for a great hackathon!**