Practice with fastText

-Requirements: Python 3.5, Gensim

-Get trained models or train your own models with data provided (EMEA corpus)

Activity 1 - Messing around with character n-grams

-The python script (`fasttext-nearest-neighb.py`) obtains the 10 most similar words with regard to an input item provided through command line.

If the word is in the vocabulary, a word vector is obtained through the word2vec-compatible file (.vec extension). This file only contains: 1) The vocabulary (words and counts); and 2) The vectors for in-vocab words. To load the .vec file, we use the following gensim method:

```
model_word2vec = gensim.models.Word2Vec.load_word2vec_format(file.vec)
```

If we test an out-of-vocabulary word, we use the fastText .bin model containing also extra information such as vectors of n-grams (if they are computed). This model is loaded through the following gensim method:

```
model_fasttext = fasttext.load_model(fasttext_model.bin)
```

*See more details in: https://github.com/RaRe-Technologies/gensim/issues/814

Act. 1.1. Exploring the effect of character n-grams

1.1.1. Load a model without character n grams :

EMEA_part_fr_ws10_0chgr_dim100_neg10
EMEA_part_fr_ws1_0chgr_dim100_neg10

Data is a subset of the French corpus from the European Medicines Agency (>7M tokens). The full data can be get at: http://opus.lingfil.uu.se/EMEA.php
Both models were trained with minimum word count of 1, vector dimension of 100, negative sampling of 10, and learning rate of 0.05 (default value in fastText).

-First test with words that might be found in the vocabulary: e.g. *aspirine, hypertension, hépatique, distal.* What did you get?

e.g. ws 1: EMEA_part_fr_ws1_0chgr_dim100_neg10

hypertension	hépatique
('portale', 0.6471356153488159)	('rénale', 0.7632142305374146)
('bradyarythmie', 0.6354788541793823)	('fonctionhépatique', 0.7263727188110352)
('lymphoedème', 0.6261867880821228)	('insuffisance', 0.7228057384490967)
('cardiopathie', 0.6235842704772949)	('multi-viscérale', 0.6862286329269409)
('elevée', 0.6209417581558228)	('multiviscérale', 0.6849429607391357)
('sibilants', 0.6200442910194397)	('légere', 0.6769933700561523)
('angiospasme', 0.6180312633514404)	('surrénalienne', 0.6605823636054993)
('tachyarythmie', 0.6129104495048523)	('stéatose', 0.6571006774902344)
('adéquatement', 0.6122598648071289)	('quadruplée', 0.6570712327957153)
('mucosités', 0.605185866355896)	('microalbuminerie', 0.6540659666061401)

-Now test a word that might be missing from the vocabulary, e.g. spelling mistakes (**adpirine*), commercial names (e.g. *nintédanib*). What did you get?

adpirine	nintédanib
Word not in corpus!	Word not in corpus!
('crête', 0.0)	('crête', 0.0)
('distal', 0.0)	('distal', 0.0)
('22001111', 0.0)	('22001111', 0.0)
('36-1-487-4100', 0.0)	('36-1-487-4100', 0.0)
('incohérent', 0.0)	('incohérent', 0.0)
('n=82', 0.0)	('n=82', 0.0)
('n=97', 0.0)	('n=97', 0.0)
('dox', 0.0)	('dox', 0.0)
('fenetre', 0.0)	('fenetre', 0.0)
('hydroxide', 0.0)	('hydroxide', 0.0)

-Try again with another model with a different window size than that tested before. Did you get a different output?

1.1.2. Load any of the following models trained with character n-grams (min.: 4; max.: 8) and the same configuration as the previous ones (except window size):

EMEA_part_fr_ws10_4-8chgr_dim100_neg10
EMEA_part_fr_ws1_4-8chgr_dim100_neg10

Try out-of-vocabulary words and compare the output:

a) EMEA_part_fr_ws1_4-8chgr_dim100_neg10

adpirine	colioscopie
Word not in corpus!	Word not in corpus!
('nélabarine', 0.9071245193481445)	('hystéroscopie', 0.8570451140403748)
('aspirine', 0.8971298933029175)	('endoscopie', 0.8485429286956787)
('vigabatrine', 0.8913751244544983)	('laparoscopique', 0.843468427658081)
('énoxoparine', 0.8813462257385254)	('macroscopique', 0.8307432532310486)
('céfalexine', 0.8790168166160583)	('endoscopique', 0.8276355266571045)
('sérine', 0.8669635057449341)	('stéréoscopique', 0.8244278430938721)
('luméfantrine', 0.8627037405967712)	('microscopique', 0.7993435859680176)
('cyclopsporine', 0.8623736500740051)	('microscopie', 0.7806546688079834)
('cyclodextrine', 0.8606017231941223)	('amblyopie', 0.7801076173782349)
('tolterodine', 0.859542965888977)('hydroxide', 0.0)	('neuroectodermale', 0.7738866209983826)

b) EMEA_part_fr_ws10_4-8chgr_dim100_neg10

adpirine	colioscopie
Word not in corpus!	Word not in corpus!
('aspirine', 0.847501277923584)	('hystéroscopie', 0.8211380243301392)
('nélabarine', 0.743427038192749)	('laparoscopique', 0.7773348093032837)
('énoxoparine', 0.7414884567260742)	('endoscopie', 0.7747243046760559)
('tolterodine', 0.7283288836479187)	('stéréoscopique', 0.7582247853279114)
('cyclopsporine', 0.7259498238563538)	('endoscopique', 0.7557596564292908)
('mépéridine', 0.7258063554763794)	('spectroscopie', 0.7504290342330933)
('fenfluramine', 0.7254996299743652)	('macroscopique', 0.7492125630378723)
('azapropazone', 0.7200309038162231)	('biomicroscope', 0.7345836162567139)
('serine', 0.7171530723571777)	('microscopie', 0.7295613884925842)
('diamine', 0.7170047760009766)	('chromatopsie', 0.7169327735900879)

Act. 1.2. Exploring the effect of window size

Load any of the previous models and compare it with the corresponding pair with a different window size. Try only in-vocabulary words. Which types of words do you get? Notice the POS-categories, derivation variants, acronyms or synonyns.

EMEA_part_fr_ws10_4-8chgr_dim100_neg10	EMEA_part_fr_ws1_4-8chgr_dim100_neg10
diabète ('insulinodépendant', 0.7695283889770508) ('non-insulinodépendant', 0.7610112428665161) ('insulinodépendants', 0.7492904663085938) ('dnid', 0.6998475790023804) ('insulino-dépendant', 0.6991622447967529) ('linodépendant', 0.6676045060157776) ('insulino-dépendants', 0.6640761494636536) ('insulino', 0.6424091458320618) ('acidocétosique', 0.6210429668426514) ('sucré', 0.6174377799034119)	diabète ('diabètes', 0.8689531087875366) ('diabétiques', 0.6818406581878662) ('diazo', 0.6552178859710693) ('diazepam', 0.6511657238006592) ('diabetes', 0.6501244306564331) ('dnid', 0.636345624923706) ('mps', 0.629019558429718) ('mpsi', 0.624091625213623) ('diabétologie', 0.622661828994751) ('non-diabétiques', 0.6191767454147339)
distal ('distale', 0.8618010878562927) ('dista', 0.8556492328643799) ('distales', 0.7538104057312012) ('nvc', 0.6558833718299866) ('distant', 0.6457308530807495) ('l2-l4', 0.6379613876342773) ('mtss', 0.6267622709274292) ('tss', 0.6263265013694763) ('interphalangienne', 0.6243892908096313) ('distance', 0.6235264539718628)	distal ('dista', 0.8699356913566589) ('distale', 0.773292601108551) ('distant', 0.6985895037651062) ('dissipé', 0.6915884613990784) ('distribué', 0.6771892309188843) ('distales', 0.6604723334312439) ('distribue', 0.6595785021781921) ('distribués', 0.658524751663208) ('distribuées', 0.648910641670227) ('distribuée', 0.6468561291694641)
hypertension ('hyperten', 0.7789657115936279) ('hypertensifs', 0.7526216506958008) ('artérielle', 0.7385604381561279) ('rénovasculaire', 0.6617308259010315) ('pression', 0.6366103887557983) ('artérielles', 0.6232606768608093) ('hypertensives', 0.6150373220443726) ('arterielle', 0.6088970899581909) ('vagale', 0.6045610904693604) ('diazo', 0.6043112277984619)	hypertension ('hypertensifs', 0.89131760597229) ('hyperten', 0.8464173078536987) ('hypertensives', 0.8154748678207397) ('hypertensive', 0.7981333136558533) ('hypertenseurs', 0.7256743311882019) ('rétension', 0.7012649774551392) ('artérielle', 0.6968148946762085) ('artériel', 0.6815078258514404) ('tension', 0.6658823490142822) ('artériels', 0.6648848056793213)

Other remarks

So far we have only tested 2 training parameters. When working with a big corpus, the parameter of minimum word count might be increased to rule out very low frequency items such as spelling mistakes, tokenization problems or hapax (rare words occurring 1 time). See the following example:

hépatique
('**fonctionhépatique'**, 0.8357903957366943)
('insuffisance', 0.7732942700386047)
('hépatiques', 0.7187262773513794)
('child-pugh-turcotte', 0.7110148668289185)
('extra-hépatique', 0.7055633664131165)
('child-', 0.7019727230072021)
('child', 0.6926040649414062)
('child-pugh', 0.6820942163467407)
('intra-hépatique', 0.6761462688446045)
('entérohépatique', 0.6750337481498718)

Activity 2. Using fastText in a supervised setting

We will use the same set of data from the EMEA corpus, but enriched with labels describing each sentence. fastText labels bear the format `__label__` + descriptor:

```
__label__treatment mélange des insulines
```

The following types of contents were used:

- medical_condition: diseases, symptoms, allergies, presence of bacteria/virus
- surgery: surgery procedures and anesthesias
- addiction: alcoholic beverages/habits, toxic drugs, smoking products/habits
- analysis: medical checkups, examinations and laboratory analysis
- treatment: medicines, treatments, vaccins, transfusions or healthcare activity
- physiol: physiological functions
- medic_dev: medical devices
- gynec_obstetr: gynecological or obstetric events
- empty: absence of any of the previous types of topics

Sentences may content different types of topics, each preceded with the `__label__` notation. For example:

```
__label__medical_condition __label__treatment diabète sucré nécessitant un traitement
```

/!\ Note that the data were annotated automatically and no revision was done /!\

2.1. Train the data

Get the EMEA annotated data.

Full data was split into 2/3 for training and 1/3 for the test:

- EMEA_part_fr_norm_lbl _ft.trn (train, 5216788 tokens)
- EMEA_part_fr_norm_lbl_ft.tst (test, 2622677 tokens)

Train using default parameters :

```
./fasttext supervised -input EMEA_part_fr_norm_lbl.trn -output
model_EMEA_superv
Read 5M words
Number of words:  30377
Number of labels: 10
Progress: 100.0%  words/sec/thread: 2259793  lr: 0.000000  loss:
0.350939  eta: 0h0m
```

2.2. Test the model

Test the model with the following command (results given in precision and recall; by default, k=1, i.e. precision and recall @1):

```
./fasttext test model_EMEA_superv.bin EMEA_part_fr_norm_u8_nca_ft.tst
P@1: 0.957
R@1: 0.728
Number of examples: 166472
```

(this represents roughly an F1-measure of 0.827).

You can also test interactively with the following command (don't forget the "-"):

```
./fasttext predict test_model_EMEA_superv.bin -

je prends de l' aspirine
__label__treatment
l' aspirine me produit des vomissements
__label__medical_condition_treatment
j'ai eu une opération
__label__surgery
je porte de lentilles
__label__medic_dev
```

2.3. Train and test other models and parameters

Other parameters are worth testing: window size, number of word/character n-grams, minimum word count, learning rate, negative sampling, number of epochs… Very sophisticated models may take longer to train.