Text mining assignment - Topic modeling

Guillem Bagaria

Objective To analyze the results output and performance from two different Topic Modeling algorithms, one algebraic (NMF) and one probabilistic (LDA).

Non-negative Matrix Factorization is a linear algebraic model for bag-of-words topic modeling, which require large and sparse matrices and are a hard problem to tackle both computationally and statistically that may cause unstable feature selection.

The goal is to find two more tractable, non-negative matrices, obtained through numeric approximation, such that their product results in the original non-negative matrix.

$$W_{m\times k} \times H_{k\times n} \approx A_{m\times n}$$

Both W and H are initialized with random, non-negative values and are iteratively updated up until convergence within a fixed threshold.

$$H_{ij} \leftarrow Hij\frac{(W^T A)_{ij}}{(W^T W H)_{ij} + \epsilon}$$

$$W_{ij} \leftarrow Wij\frac{(AH^T)_{ij}}{(WHH^T)_{ij} + \epsilon}$$

Latent Dirichlet Allocation is a probabilistic model that finds the most likely membership of a document from k number of topics. The learning process uses variational Bayes in order to update the distribution of the likelihood for each word and thus, document.

One of the expected results is that LDA would perform better for big data. Time complexity is polynomial in NMF while In LDA, the time complexity is lower, proportional to $\mathcal{O}(n*\text{iterations})$ where n is the number of documents.

The data This dataset is standard in the SciKitLearn python package and contains around 1800 newsgroup posts from 20 threads, which correspond to the 20 different topic labels. For this application, only the first level group will be used, in order to simplify the analysis.

Default topics ’alt.atheism’, ’comp.graphics’, ’comp.os.ms-windows.misc’, ’comp.sys.ibm.pc.hardware’, ’comp.sys.mac.hardware’, ’comp.windows.x’, ’misc.forsale’, ’rec.autos’, ’rec.motorcycles’, ’rec.sport.baseball’, ’rec.sport.hockey’, ’sci.crypt’, ’sci.electronics’, ’sci.med’, ’sci.space’, ’soc.religion.christian’, ’talk.politics.guns’, ’talk.politics.mideast’,’talk.politics.misc’, ’talk.religion.misc’

Target topics Atheism, Christianity, Computer Hardware, Windows OS, For Sale, Motor Sports, Team Sports, Cryptography, Electronics, Medicine, Space, Gun Politics, Middle East Politics, Other Politics.

The raw text from the posts is then processed and appropriately tokenized. As this is an unsupervised learning approach, all the data will be used for model training.

Results Findings of the comparison between different algorithms and settings. The selected number of topics was reduced from the original specific 20 to 14 more general categories for each of the topics in order to produce slightly more manageable results while remaining sensible.

LDA tf

```
Topic 0:
people armenian said israel armenians
war jews turkish men israeli
Topic 1:
space drive disk hard data scsi power
drives speed controller
Topic 2:
time good case don used use like book
does actually
Topic 3:
key chip encryption keys clipper public
use security privacy technology
Topic 4:
new vs bike la san period st york
chicago pts
Topic 5:
edu com file mail available ftp information
db send list
Topic 6:
don just like think know people going
time ve good
Topic 7:
game team games year season play hockey
players win league
Topic 8:
law people government gun state right
rights control guns crime
Topic 9:
windows use program window software
file dos using output image
Topic 10:
god people jesus believe does say
think christian bible don
Topic 11:
ax max end air information drivers
video conference new use
Topic 12:
thanks know does use problem help work
like need ve
Topic 13:
mr president university research health
april national center program stephanopoulos
```

NMF tf-idf

```
Topic 0:
don think good time ve really did make
way want
Topic 1:
thanks mail advance looking hi info
email information address help
Topic 2:
god jesus bible christ believe faith
christian christians church lord
Topic 3:
game team games year season players
play hockey win league
Topic 4:
drive scsi drives hard disk ide floppy
controller cd tape
Topic 5:
car cars engine miles price new speed
condition good bike
Topic 6:
windows file files dos use window
program using problem running
Topic 7:
key chip encryption clipper keys escrow
government use public algorithm
Topic 8:
people government israel armenian
israeli jews armenians rights state
law
Topic 9:
edu soon cs com email university ftp
send internet david
Topic 10:
card video monitor cards drivers bus
vga sale color driver
Topic 11:
just wondering thought don wanted mean
tell fine oh maybe
Topic 12:
does know anybody don let mean doesn
help program appreciated
Topic 13:
like sounds looks look sound lot things
new doing sell
```

NMF tf In this case the matrix is fed with simple term frequency:

```
Topic 0:
ax max end air follow firearms fit
floppy folks young
Topic 1:
edu com available graphics ftp pub
image mail data send
Topic 2:
db cs al bits higher gas lower bit
west east
Topic 3:
people said know don didn just like
went say think
Topic 4:
file gun control firearms states united
mr house law crime
Topic 5:
mr stephanopoulos president know going
don think said did package
Topic 6:
jpeg image gif file color images format
quality version files
Topic 7:
output file entry program stream build
rules section info line
Topic 8:
hockey league new nhl team season edu
games vs division
Topic 9:
internet privacy anonymous information
email use mail computer pub electronic
Topic 10:
use widget window subject application
available xt motif set used
Topic 11:
disk drive drives hard bios rom controller
card floppy supports
Topic 12:
space launch satellite new data nasa
program commercial south year
Topic 13:
god jesus does people atheists believe
atheism bible religious good
```

NMF tf-idf In this case the matrix uses term frequency inverse document frequency:

```
Topic 0:
don think good time ve really did make
way want
Topic 1:
thanks mail advance looking hi info
email information address help
Topic 2:
god jesus bible christ believe faith
christian christians church lord
Topic 3:
game team games year season players
play hockey win league
Topic 4:
drive scsi drives hard disk ide floppy
controller cd tape
Topic 5:
car cars engine miles price new speed
condition good bike
Topic 6:
windows file files dos use window
program using problem running
Topic 7:
key chip encryption clipper keys escrow
government use public algorithm
Topic 8:
people government israel armenian
israeli jews armenians rights state
law
Topic 9:
edu soon cs com email university ftp
send internet david
Topic 10:
card video monitor cards drivers bus
vga sale color driver
Topic 11:
just wondering thought don wanted mean
tell fine oh maybe
Topic 12:
does know anybody don let mean doesn
help program appreciated
Topic 13:
like sounds looks look sound lot things
new doing sell
```

Conclusions

Initial observations show that NMF presents coherent topic words, especially when combined with the tf-idf input matrix. Simple tf performs makes some serious mistakes, such as missing the "Middle East Politics" and misplacing many words; Topic 0 (firearms, floppy [disk]), Topic 2 (bits, gas), etc.

Although the results for NMF (tf-idf) and LDA (tf) are similar, the execution time of NMF however is much lower than LDA, and in certain situations might prove to be a determining factor in its use.

	Execution time
NMF	1.79s
LDA	18.01s

LDA produces very coherent topics even with simple term frequency data input.

Convergence time and topic modeling performance might improve in LDA when used in a semi-supervised learning scenario, with strong priors. This would be the next step.