Xinhao Liao
March 21st, 2020

SI630 Homework 4 – Latent Dirichlet Allocation

1 Implementing a Gibbs Sampler

The implementation is finished based on the skeleton code in *lda.py*.

2 Running on the Code Real Data

With the model implemented, we can run it on data in the directory *wiki* for 1000 iterations (with a vocabulary of 1000 words) to obtain 5 topics, and measure the needed on PowerShell with the command

```
Measure-Command {python lda.py --doc_dir ./wiki --output res --num_iterations 1000}
```

which gives

```
Days              : 0
Hours             : 0
Minutes           : 3
Seconds           : 17
Milliseconds      : 256
Ticks             : 1972560538
TotalDays         : 0.00228305617824074
TotalHours        : 0.0547933482777778
TotalMinutes      : 3.28760089666667
TotalSeconds      : 197.2560538
TotalMilliseconds : 197256.0538
```

showing that it takes 3 minutes 17 seconds and 256 milliseconds.

The result with 50 most probable words per topic is given in *res.topic*, and I rename the file to be *my-topics.txt* as required. (Note: method *report* of class *LdaTopicCounts* in *lda.py* is reimplemented to generate words in the order of word-in-topic probabilities.)

3 Comparing models

To compare with implementations of Gensim and Mallet, I coded *lda_gensim.py* and *lda_mallet.py* which applies the two implementation on the wiki articles 1000 iterations to generate 5 topics and give 50 most probable words per topic.

I tested Gensim's implementation with the command

```
Measure-Command {python lda_gensim.py}
```

which gives

```
Days              : 0
Hours             : 0
Minutes           : 0
Seconds           : 14
Milliseconds      : 433
```

```
Ticks             : 144330438
TotalDays         : 0.000167049118055556
TotalHours        : 0.00400917883333333
TotalMinutes      : 0.24055073
TotalSeconds      : 14.4330438
TotalMilliseconds : 14433.0438
```

showing that it takes only 14 seconds and 433 milliseconds, as well as some output of the training process. The result with 5 topics and 50 most probable words per topic is written in *gensim-topics.txt*.

Similarly, I tested Mallet's implementation with the command

```
Measure-Command {python lda_mallet.py}
```

which gives

```
Days              : 0
Hours             : 0
Minutes           : 0
Seconds           : 39
Milliseconds      : 920
Ticks             : 399205804
TotalDays         : 0.00046204375462963
TotalHours        : 0.0110890501111111
TotalMinutes      : 0.665343006666667
TotalSeconds      : 39.9205804
TotalMilliseconds : 39920.5804
```

showing that it takes 39 seconds and 920 milliseconds. The result with 5 topics and 50 most probable words per topic is written in *mallet-topics.txt*.

3.1 Whether each model found reasonable topics?

3.1.1 My implementation

The results of my implementation is

```
------------
Topic 0
------------
one, also, used, two, many, time, known, well, another, due, number, long, include,
could, air, different, series, several, large, made, found, four, back, point, old,
make, early, based, high, though, way, taken, modern, world, name, took, three, show,
model, new, open, n't, never, letters, book, set, occurs, seen, original, followed
------------
Topic 1
------------
village, white, county, species, st., south, family, new, also, road, central, built,
church, known, park, roman, per, name, parish, average, regional, two, eastern, light,
missouri, york, major, near, approximately, large, australia, victoria, act, total,
bishop, served, bath, area., base, communities, via, described, mary, former, club,
ranges, made, famous, metres, found
------------
Topic 2
------------
game, forest, team, national, played, games, service, professional, cup, play, club,
```

```
second, world, rugby, players, college, baseball, round, points, championship, winning,
major, race, runs, competition, 2012, speedway, 2002, championships, top, scored,
chicago, figure, competed, track, moved, tournament, basketball, skating, olympics,
team., nation, jr., teams, division, rounds, motor, position, office, 1994
------------
Topic 3
------------
new, also, would, born, school, university, national, member, two, one, october,
september, life, best, august, high, made, november, house, second, reported, time,
joined, served, received, currently, many, early, international, home, 2008, several,
three, began, january, john, 2007, four, july, march, place, association, california,
2006, former, day, started, political, 2005, moved
------------
Topic 4
------------
camp, training, music, stewart, musical, john, love, song, robert, world, played,
theatre, featured, ace, recorded, father, danny, shows, elizabeth, charlie, book,
san, production, division, mother, cast, assigned, ruth, hall, design, directed, tony,
flying, scott, famous, series, lasted, texas, lyrics, greg, knew, hill, starred,
louis, pearl, twelve, play, festival, renamed, success
```

Among them, I find that Topic 2 is the most clear. With words like game, cup, club, rugby, baseball, championship, race, tournament, basketball, skating, olympics, and teams shown together, we can easily think of the topic about sports.

It also makes sense to find that a lot of names like stewart, john, robert, danny, elizabeth, charlie, tony, scott, louis, along with words like music, musical, theatre, father, mother, famous, lyrics, starred, success clustered together in Topic 4. We may reasonably assume that this is a topic about biographies of some famous figures.

Also, years like 2005, 2006, 2007, 2008, and months like january, march, july, august, september, october, november are all given in topic 3. It might represent a topic about organizations given that words like school, university, national, member, joined, international, began, association, and political are also related to the topic.

As for other topics, though we can still find some underlying logic, it seems not so clear as topic 2, 3 and 4.

3.1.2 Gensim's implementation

The topics and words given by Gensim's implementation are

```
------------
topic 0
------------
also, reported, one, october, september, air, new, many, national, political, two,
day, would, san, known, met, three, time, stated, house, south, led, member, could,
win, early, world, due, large, currently, joined, alliance, china, made, several,
used, another, march, meet, making, seen, 2007, meeting, four, event, moved, central,
total, late, former
------------
topic 1
------------
one, new, village, would, two, made, world, time, born, november, family, best, also,
series, story, york, several, found, high, battle, king, include, july, central, many,
former, long, october, known, life, old, major, son, well, four, ice, white, eastern,
```

```
played, took, south, back, joined, english, member, january, school, though,
early, house
------------
topic 2
------------
also, university, game, white, one, national, school, member, john, two, served,
born, would, place, house, played, known, new, association, roman, competition, many,
best, name, musical, team, made, second, used, st., time, home, missouri, family, games,
received, theatre, act, currently, son, world, production, several, high, early, show,
young, international, county, october
------------
topic 3
------------
also, one, new, two, used, known, national, forest, time, many, service, would, school,
game, well, county, born, made, camp, world, high, team, played, four, several, early,
species, number, due, different, university, three, large, second, show, another,
series, include, training, life, member, could, california, august, best, canadian,
long, moved, found, based
------------
topic 4
------------
also, many, letters, park, new, used, one, national, two, second, points, born, team,
2002, south, speedway, race, built, men, school, made, standard, began, john, four,
st., figure, time, several, former, motor, racetrack, drivers, track, 1993, would,
skating, ontario, high, coliseum, august, member, took, day, name, top, racing,
cup, sprint, jr.
```

Though it makes sense to sometimes, for example to have speedwat, race, motor, racetrack, drivers, racing in one topic, the overall topic themes are not so coherent.

3.1.3 Mallet's implementation

The topics and words given by Mallet's implementation are

```
------------
topic 0
------------
game, team, school, played, games, national, club, professional, high, world, college,
cup, place, competition, play, players, rugby, born, 2002, winning, baseball,
championship, moved, runs, show, 2008, 2005, round, 2007, figure, championships,
sports, basketball, hall, competed, time, association, county, top, kansas, senior,
led, tournament, home, skating, men, olympics, team., allowed, win,
------------
topic 1
------------
born, university, member, october, september, national, served, camp, made, november,
house, reported, early, received, york, school, began, joined, life, san, march, august,
john, training, california, july, january, international, son, position, service,
family, south, father, young, home, february, st., roman, association, met, political,
division, day, high, moved, russian, 2012, central, late,
------------
topic 2
------------
```

```
music, stewart, musical, robert, theatre, featured, love, john, played, danny, song,
charlie, elizabeth, speedway, recorded, production, show, series, book, ace, cast,
track, tony, jr., ruth, 2012, father, race, play, greg, points, lyrics, motor, starred,
ahead, cup, tells, scott, louis, richard, include, sprint, nominated, shows, nascar,
jim, cars, johnson, chase, edwards,
------------
topic 3
------------
time, number, due, long, air, world, series, based, include, large, found, made, high,
point, led, make, back, early, set, story, model, included, show, modern, original,
book, n't, major, built, letters, top, day, occurs, battle, case, political, front,
open, shows, life, parts, individual, making, added, commonly, production, family,
pilot, time., ten,
------------
topic 4
------------
forest, village, species, national, county, white, service, canadian, canada, south,
road, parish, built, park, church, central, major, large, act, regional, family,
approximately, area., average, eastern, st., missouri, total, australia, base, bath,
office, ontario, mary, metres, communities, ranges, made, canada., victoria, run, 2006,
include, burnell, references., castle, racing, coliseum, racetrack, entire,
```

Like the results in first implementation, there seems a topic (Topic 0) about sport with words game, team, played, games, club, cup, players, rugby, baseball,championship, sports, basketball, tournament, skating, olympics, etc.

And there also seem to be a topic (Topic 2) about biographies of some famous figures, with names like stewart, robert, john, danny, charlie, elizabeth, tony, scott, louis, richard, jim, johnson, and edwards, with words like music, musical, theatre, song, lyrics, starred, nominated, etc.

For Topic 1 here, it's also quite similar to topic 3 in first implementation, with months like january, march, july, august, september, october, and november, and words like university, member, national, school, joined, life, international, political, etc shown together. But years like 2005, 2006, 2007, and 2008 do not appear this time. We might still assume that it's a topic about organizations, but the theme is no so coherent as it was in first implementation's result.

And again the logic for other topics is not so clear.

3.2 Whether each model found the same kind of topics?

As shown and analyzed above, our model and Mallet's implementation find quite similar topics, while the result of Gensim's model is much more different. One reason might be that both our and Mallet's impelmentation uses Gibbs sampling.

3.3 How the models differed in speed and whether this seems related to topic quality

As tested before, our self-completed model is the slowest, which spent 3 minutes 17 seconds and 256 milliseconds on the 1000 iterations. Gensim's implementation is the fastest, which spent only 14 seconds and 433 milliseconds. And it takes 39 seconds and 920 milliseconds for the Mallets's implemetation to finish the 1000 iterations.

As for the qualities, the result our model seems to be the best. The Mallet's model seems to be slightly worse. And the Gensim's model seems to be the worst.

Based on what is discovered above, it seems that a faster model tends to give worse topic quality.