Computational Linguistics
Assignment 5
Latent Dirichlet Allocation

Sangeet Sagar
sasa00001@stud.uni-saarland.de

January 29, 2021

1 Introduction

LDA is a generative probablistic framework that was proposed to identify topics that collectively represented a group of words in a document. LDA is based on the assumption that a document consists of a combination of various *latent* topics and each topic can be multinomial probablity density over words. Gibbs sampling method helps approximate the the posterior distribution and thus to learn the model from data [1].

This assignment implements the Gibbs sampler which resamples a topic for each word in the corpus according to the probability distribution in formula 5 [2]

2 Description

The script:

- `main.py`: this is the main script that implements the gibbs sampler which you should be running. **Extra Credit**: Entire implmentation done in `numpy` that took about 3.4 hours as total runtime in a machine of 16 G RAM and 12 Cores CPU.

3 Requirements

The scripts have been tested on:

1. Python: `3.8.3`

2. Numpy: `1.19.2`

4 Project file structure

```
├── data
├── main.py
├── README.md
├── report
│   └── ass5_report.pdf
├── results
│   ├── topicwise_words.txt
│   └── training.log
├── results2
│   ├── run_me.sh
│   ├── topicwise_words.txt
│   └── training.log
├── results3
│   ├── run_me.sh
│   ├── topicwise_words.txt
│   └── training.log
└── results4
├── run_me.sh
├── topicwise_words.txt
└── training.log
```

5 Usage

- **Help**: for instructions on how to run the script with appropriate arguments.
  `python main.py -help`

```
usage: main.py [−h]
               [−alpha ALPHA]
               [−beta BETA]
               [−num_topics NUM_TOPICS]
               [−epochs EPOCHS]
               [−num_sents NUM_SENTS]
               [−num_top_words NUM_TOP_WORDS]
               corpus_f
               out_dir

Implements a Gibbs sampler which resamples a topic for each word in the corpus
according to the probability distribution in formula [5] of
(Griffiths & Steyvers 2004)

positional arguments:
  corpus_f              path to input text corpus file
  out_dir               path to save the frequent words for each topic

optional arguments:
  −h, −−help            show this help message and exit
  −alpha ALPHA          Parameter that sets the topic distribution for the documents
  −beta BETA            Parameter that sets the topic distribution for the words
  −num_topics NUM_TOPICS
                        number of topics
  −epochs EPOCHS        number of training iterations
  −num_sents NUM_SENTS  number of sentences to train from
  −num_top_words NUM_TOP_WORDS
                        number of words to select from each topic
```

6 Datatset

Trained on the corpus of 2000 movie reviews from Pang & Lee (2004).

7 Runtime

- **Total** runtime: 3.431 hrs
- Each iteration runtime: 23 s
- Data loading runtime: 0.166 s
- Initialization runtime: 3.329 s

8 Results

- Use default parameter: `-alpha 0.02 -beta 0.1 -num_topics 20 -epochs 500 -num_top_words 10`

```
python main.py movies-pp.txt results/
```

```
Topic 1: ryan, war, hanks, private, spielberg, babe, chicken, saving, pig, red.
Topic 2: alien, aliens, planet, truman, ship, space, mars, earth, science, effects.
Topic 3: titanic, shakespeare, ship, sandler, wedding, love, cameron, angels, rose, singer.
Topic 4: dvd, grace, rocky, finn, rock, songs, disc, horse, paulie, hedwig.
Topic 5: comedy, smith, funny, ben, bob, jay, school, sex, west, football.
Topic 6: murphy, comedy, eddie, roberts, kate, romantic, cole, julia, willis, cusack.
Topic 7: war, joe, men, army, american, general, political, washington, soldiers, battle.
Topic 8: girls, flynt, evil, spice, austin, patch, powers, dr, jackal, frank.
Topic 9: star, wars, lucas, phantom, jedi, menace, effects, nbsp, contact, series.
Topic 10: nights, van, black, boogie, derek, anderson, 54, fugitive, ford, jones.
Topic 11: vampire, horror, burton, blade, vampires, house, carpenter, blair, scary, witch.
Topic 12: jackie, chan, action, kong, hong, van, damme, bond, martial, chinese.
Topic 13: tarantino, harry, crime, jackie, simon, cop, pulp, fiction, brown, jackson.
Topic 14: batman, arnold, cage, robin, max, seagal, snake, schumacher, wrestling, impact.
Topic 15: film, movie, one, like, even, good, would, time, get, much.
Topic 16: godzilla, troopers, starship, verhoeven, bulworth, broderick, wild, besson, bacon, robocop.
Topic 17: disney, family, animated, animation, mulan, children, toy, voice, kids, king.
Topic 18: scream, horror, killer, trek, 2, urban, julie, slasher, williamson, summer.
Topic 19: film, one, story, life, like, character, movie, characters, man, also.
Topic 20: tarzan, spawn, black, jane, cauldron, prinze, carry, jawbreaker, gladiator, liz.
```

Nonetheless, with given parameters set to default, indeed they ended up generating some coherent words that can be categorised into a topic. Lets try to categorise each of them:

- Topic 1: Army and war related
- Topic 2: Alien and extraterrestrial
- Topic 3: Romance
- Topic 4: Movie related and characters
- Topic 5: Comedy
- Topic 6: TV Characters
- Topic 7: Politics and war
- Topic 8: Book characters
- Topic 9: Star-Wars
- Topic 10: TV Drama
- Topic 11: Horror
- Topic 12: Action and thriller
- Topic 13: Movies
- Topic 14: –
- Topic 15: –
- Topic 16: Sci-fi
- Topic 17: Animation and Comedy
- Topic 18: Horror and Thriller
- Topic 19: –
- Topic 20: Fictional Character

- **Extra Credit**: Trying out different number of topics and different values of hyperparameters.

  1. `python main.py data/movies-pp.txt results2/`
     `-alpha 0.05 -beta 0.5 -num_topics 30 -num_top_words 20`
     For complete list of topic wise words: `cat results2/topicwise_words.txt`

```
Topic 1: broderick, franklin, ronna, diedre, fichtner, tracy, liman, stanton, pi, payne, _election_,
         colqhoun, polley, falk, _rushmore_, regiment, jadzia, olin, massachusetts, molly.
Topic 2: jakob, _the, chad, thirteenth, daisy, caveman, stargher, valentine_, jasmine, ulee, romulus,
         jermaine, ghetto, andreas, fuller, hartman, caligula, catharine, tandy, turboman.
Topic 3: roxbury, briggs, alchemy, louisa, intimacy, eastwood, goop, cristoffer, nell, rylance, dietz,
         greenfingers, tran, grieco, hung, routines, everett, beacham, greenleaf, spall.
Topic 4: derek, rudy, francis, dead-bang, skinhead, zoolander, x, zach, deceiver, kaye, kings, gates,
         wayland, beck, skinheads, coyle, barlow, harmon, kennesaw, lesbos.
Topic 5: vampire, vampires, blade, carpenter, snipes, crow, wesley, lumumba, ghosts, valek, macleane,
         pam, plunkett, onegin, squad, woods, lillian, mars, peck, gia.
```

2. `python main.py data/movies-pp.txt results3/`
   `-alpha 2  -beta 0.1 -num_topics 25 -num_top_words 25`
   For complete list of topic wise words: `cat results3/topicwise_words.txt`

```
Topic 1: movie, first, people, time, show, two, new, well, years, movies, one, would, tv,
         see, home, made, film, three, watch, many, world, screen, 1, last, back.
Topic 2: james, bond, william, role, king, peter, charles, never, richard, henry, carry, gibson,
         jerry, scott, although, joan, mel, however, gone, head, played, cast, part, opening, british.
Topic 3: world, life, us, way, city, death, like, human, man, dark, reality, message, black, god,
         new, place, society, questions, people, david, live, game, something, look, camera.
Topic 4: film, one, even, almost, novel, would, characters, new, mr, hollywood, less, director,
         character, book, many, upon, years, could, effective, whose, often, story, based, despite, may.
Topic 5: ship, crew, trek, first, titanic, water, island, disaster, cameron, story, star, deep, last,
         monster, jack, boat, virus, effects, part, cast, special, next, well, members, ocean.
```

3. `python main.py data/movies-pp.txt results4/`
   `-alpha 1.67  -beta 0.1 -num_topics 30 -num_top_words 20`
   For complete list of topic wise words: `cat results4/topicwise_words.txt`

```
Topic 1: show, truman, jones, dr, evil, mike, powers, carrey, world, spawn, austin,
         island, martin, jim, tommy, ford, shrek, lee, friend, fugitive.
Topic 2: music, rock, mars, tim, mission, band, burton, apes, flynt, planet, musical,
         sequence, larry, snake, human, hollow, songs, brian, song, wife.
Topic 3: series, mr, television, show, nights, anderson, william, less, upon, x-files,
         boogie, tv, summer, roberts, nbsp, ms, fans, consider, may, screen.
Topic 4: school, high, kids, teacher, boy, student, american, girl, paul, 10, jim,
         parents, football, around, team, girlfriend, pie, boys, cole, teenage.
Topic 5: new, deep, york, days, godzilla, action, impact, effects, park, summer, special,
         arnold, disaster, stop, armageddon, monster, world, schwarzenegger, team, end
```

References

[1] William M. Darling. A theoretical and practical implementation tutorial on topic modeling and gibbs sampling. 2011.

[2] Thomas L. Griffiths and Mark Steyvers. Finding scientific topics. *Proceedings of the National Academy of Sciences*, 101(suppl 1):5228–5235, 2004.