2nd Year Project – Milestone 1: Lyrics era prediction

Barbara Plank

February 8, 2019

Hey man, every-every-everybody's talkin' about it, everybody's talkin' bout...

Guess: Which era did this song lyric come from? Possible labels are: $y \in \{$pre1980s, 1980s, 1990s, 2000s$\}$, where 2000s is anything from 2000 or later.

In this milestone you will build a system that can automatically classify song lyrics by era. You will:

- Do basic text processing, tokenizing your input and converting it into a bag-of-words representation
- Build a machine learning classifier based on the generative model, using Naive Bayes
- Build a machine learning classifier based on the discriminative model, using a logistic regression classifier implemented in Keras
- Evaluate your classifiers and examine what they have learned
- Implement techniques to improve your classifier. Who will build the best classifier on a held-out test set? Submit your best prediction file.

Requested reading: Chapters 4 and 5 of [1]

1 Setup

You will need the following packages, besides python 3.6 in Anaconda

1. jupyter
2. numpy
3. matplotlib
4. pandas

5. nose
6. keras
7. tensorflow

We suggest you create a fork of the course repository in your git repository and keep track of your work there. You can start solving your milestone on your laptop, but it will soon come in handy to use the dedicated computing machine to run the code. You can do so by connecting to your dedicated group server (as explained in the lab material) and keeping a copy of your git repository there.

2 Milestone 1

Solve the exercises provided in the notebook `milestone1.ipynb`.

The outcome of your milestone are three files: one slide (as pdf) with your a representative finding of this milestone (e.g., a result graph or similar, you can be creative here; imagine you need to show/summarize your milestone in a single slide, in general less text is more), a prediction file on held-out data, and a link to your github repository.

Submit Replace dsproj01 with your own username and:

pdf Submit a pdf of your one slide summary of milestone 1. Name it: `dsproj01-m1-slide.pdf`

txt Submit your prediction file on the held out test data (this will be used to rank your system's performance against those of your fellow peers). Name it: `dsproj01-m1-predictions.txt`

git Submit a link to your git repository where your notebook is with the solutions. The link should be stored in this file: `dsproj01-m1-git.txt`

Submission deadline: The submission deadline for this exercise is **Monday, February 18** at 14:00 CET.

Submission instructions: Upload your pdf and prediction file to the course repository github page, i.e.,

```
git clone https://github.itu.dk/bapl/2ndyearproject-2019.git
(or go to the folder where you have cloned this repo before
    and run git pull origin master)

cd 2ndyearproject-2019
cd milestone1
```

Copy your three files into the folder and upload them, using your assigned user name. Make sure you follow the following naming convention, assuming you are user dsproj01:

```
git add dsproj01-m1-predictions.txt
git add dsproj01-m1-slide.pdf
git add dsproj01-m1-git.txt
git commit -a
git push origin master
```

References

[1] Jurfasky and Martin, In Preparation. *Speech and Language Processing (3rd ed. draft)*. Available at https://web.stanford.edu/~jurafsky/slp3/