Assignment 1 - Corpus Creation

Marko Ložajić

University of Tübingen

June 5, 2019

General remarks

- Great job!

General remarks

- Great job!
- Please include honor code

General remarks

- Great job!
- Please include honor code
- Please do not commit to assignment repository after deadline

General remarks

- Great job!
- Please include honor code
- Please do not commit to assignment repository after deadline
- Reminder: worst lab doesn't count!

General (Python) comment

```
# Recommended way to open a file in Python 3:
with open(file, "r") as f:
    ...
```

```
# As opposed to:
f = open(file, "r")
...
f.close()
```

```
# As well as:
import io
with io.open(file, "r") as f:
    ...
```

Recap

1. Construct distinctive word lists for English and German
2. Collect tweets containing words in both languages
3. Have `langdetect` tell you how you did

Part 1 - Getting the most frequent words (sketch)

Goal: We want the 5000 most frequent English words and the most frequent 2000 German words, such that no words within the most common 20000 of the other language are selected.

```
import gzip
from collections import Counter

with gzip.open(corpus, "rt", encoding="utf-8") as f:
    c = Counter()
    for line in f:
        for word in line.split():
            if len(word) > 3:
                c[word] += 1  # or word.lower()
    # return all words, not just first 20,000!
    return [word for word, _freq in c.most_common()]
```

Part 1 - Saving word lists (sketch)

```
common_words = set(en_words[:20000]) \
               & set(de_words[:20000])
en_wordlist = []
collected = 0
for word in en_words:
    if word not in common_words:
        en_wordlist.append(word)
        collected += 1
        if collected == 5000:  # 2000 for German
            ...  # write words to file and break
```

Part 1 - Saving word lists (Option 2)

```
en_wordlist = set(en_words[:5000])
de_wordlist = set(de_words[:2000])
intersec = eng_wordlist & de_wordlist
words_discarded = 0

while len(intersec) > 0:
    en_wordlist -= intersec
    de_wordlist -= intersec
    en_wordlist.add(set(
        eng_words[5000 + words_discarded:
                  5000 + words_discarded + len(intersec)]))
    de_wordlist.add(set(
        de_words[2000 + words_discarded:
                 2000 + words_discarded + len(intersec)]))
    words_discarded += len(intersec)
    intersec = en_wordlist & de_wordlist

... # write word lists to files
```

Part 2 - Getting code-switched tweets

- Search API vs Stream API

Part 2 - Getting code-switched tweets

- Search API vs Stream API
- Fighting rate limits

Part 2 - Getting code-switched tweets

- Search API vs Stream API
- Fighting rate limits
- Fetching unique tweets

Part 2 - Getting code-switched tweets

- Search API vs Stream API
- Fighting rate limits
- Fetching unique tweets
- Speeding up tweet hunt

Part 2 - Getting code-switched tweets

Some concrete possibilities for the search API:

- query? "-filter:retweets", part of German word list

Part 2 - Getting code-switched tweets

Some concrete possibilities for the search API:

- query? "-filter:retweets", part of German word list
- count? 100

Part 2 - Getting code-switched tweets

Some concrete possibilities for the search API:

- query? "-filter:retweets", part of German word list
- count? 100
- max_id?

Part 2 - Getting code-switched tweets

Some concrete possibilities for the search API:

- query? "-filter:retweets", part of German word list
- count? 100
- max_id?
- geocode? "48.6,11.5,400km", list of countries

Part 2 - Getting code-switched tweets

Some concrete possibilities for the search API:

- query? "-filter:retweets", part of German word list
- count? 100
- max_id?
- geocode? "48.6,11.5,400km", list of countries
- tweet_mode? "extended"

Part 2 - Getting code-switched tweets (sketch)

```
en_wordset = read_words("en.words")
de_wordset = read_words("de.words")
for tweet in tweets:
    if len(tweet.full_text < 50):
        continue
    en, de = False, False
    for word in tweet:
        if word in en_wordset:
            en = True
        elif word in de_wordset:
            de = True
    if en and de:
        ... # dump to file, check if 50 tweets reached
```

Gallery

- Like Italy is a 10 und Schweiz naja 7 oder so. Still high aber like nichts gegen Italien bitte

Gallery

- Like Italy is a 10 und Schweiz naja 7 oder so. Still high aber like nichts gegen Italien bitte
- Die midlife crisis des millennial startet bei der Überlegung sich beim poetry slam anzumelden

Gallery

- Like Italy is a 10 und Schweiz naja 7 oder so. Still high aber like nichts gegen Italien bitte
- Die midlife crisis des millennial startet bei der Überlegung sich beim poetry slam anzumelden
- Excuse me, wieso steht dieser komplett creepy aussehende Bär da neben dem Altar?!

Gallery

- Like Italy is a 10 und Schweiz naja 7 oder so. Still high aber like nichts gegen Italien bitte
- Die midlife crisis des millennial startet bei der Überlegung sich beim poetry slam anzumelden
- Excuse me, wieso steht dieser komplett creepy aussehende Bär da neben dem Altar?!
- Lots of sports ball fans excited for their sports ball and lots of polizei should they get too excited.

Gallery...?

- Erste Vertragsgespräche werden nach dem Pokalspiel gegen Worms geführt.

Gallery...?

- Erste Vertragsgespräche werden nach dem Pokalspiel gegen Worms geführt.
- Ok, dann muss ich halt Bill Gates werden, um von älteren angesehen zu werden lol

Gallery...?

- Erste Vertragsgespräche werden nach dem Pokalspiel gegen Worms geführt.
- Ok, dann muss ich halt Bill Gates werden, um von älteren angesehen zu werden lol
- OMG wasn los??? Leider von hier aus keine Hilfe möglich...

Gallery...?

- Erste Vertragsgespräche werden nach dem Pokalspiel gegen Worms geführt.
- Ok, dann muss ich halt Bill Gates werden, um von älteren angesehen zu werden lol
- OMG wasn los??? Leider von hier aus keine Hilfe möglich...
- Ja abwarten aber das ist der perfekte Mann für die defensive gewesen schon mal.

Gallery...?

- Erste Vertragsgespräche werden nach dem Pokalspiel gegen Worms geführt.
- Ok, dann muss ich halt Bill Gates werden, um von älteren angesehen zu werden lol
- OMG wasn los??? Leider von hier aus keine Hilfe möglich...
- Ja abwarten aber das ist der perfekte Mann für die defensive gewesen schon mal.
- I had told Liz Johnston that their tote bag was going places

Gallery...?

- Erste Vertragsgespräche werden nach dem Pokalspiel gegen Worms geführt.
- Ok, dann muss ich halt Bill Gates werden, um von älteren angesehen zu werden lol
- OMG wasn los??? Leider von hier aus keine Hilfe möglich...
- Ja abwarten aber das ist der perfekte Mann für die defensive gewesen schon mal.
- I had told Liz Johnston that their tote bag was going places
- Bei mir hats auch gedauert..

Gallery...?

- Erste Vertragsgespräche werden nach dem Pokalspiel gegen Worms geführt.
- Ok, dann muss ich halt Bill Gates werden, um von älteren angesehen zu werden lol
- OMG wasn los??? Leider von hier aus keine Hilfe möglich...
- Ja abwarten aber das ist der perfekte Mann für die defensive gewesen schon mal.
- I had told Liz Johnston that their tote bag was going places
- Bei mir hats auch gedauert..
- Jamie si tu vois ce tweet sache que je fais ça juste pour toi

Part 3 - Using langdetect

- Read in JSON file containing tweets
- Tokenize tweets using TweetTokenizer
- Report number of code-switched tweets
- Save all tokens with corresponding languages to file

Part 3 - Saving langdetect results (sketch)

```
for tweet in tweets:
    text = tweet_tokenizer.tokenize(tweet['full_text'])
    en, de = False, False
    for token in text:
        f.write(token + "\t")
        if not token.isalpha():
            f.write("OTHER\n")
            continue
        detected_languages = detect_langs(token)
        for i, lang in enumerate(detect_langs(token)):
            if lang.lang == "en":
                en = True
                f.write("en\n")
                break
            elif lang.lang == "de":
                ... # analogous to above
            elif i == len(detected_languages) - 1:
                f.write("OTHER\n")
    if en and de: code_switched_count += 1
```

Haben Sie questions?