**Subject:** Re: Meeting catch up
**Date:** Friday, 14 February 2020 at 12.09.44 Central European Standard Time
**From:** Morena Rivato
**To:** Rebekah Brita Baglini
**CC:** Marco Hubert
**Attachments:** image001.png, image002.png, image003.png, Dictionaries.xlsx, Topics.xlsx

Hi Rebekah,

I had the chance to talk with the IT and the server manager. Unfortunately, there is some reluctance in sharing the server outside the department.
I will investigate this further in order to understand what can be done.

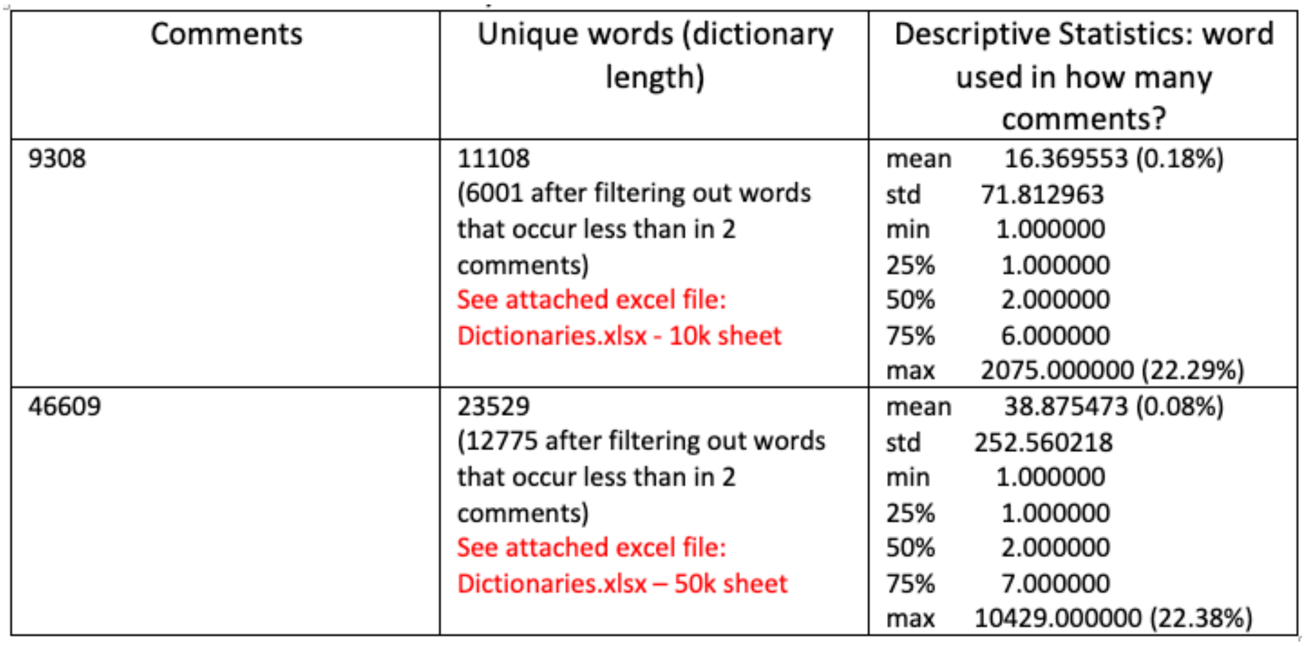
Meanwhile, below there is the current status of the project based on what we talk about. Hope it makes sense and if you have something to add/change based on your note feel free to do so.

I aim to get the new data on Monday. So I can share with you these, walk you through the current code and discuss what can be done to improve it and what possible paths we can take..

My next week it's all free, let me know if you're available for a meeting. I can reach you there.

**CURRENTLY NOT WORKING:**

- We are dealing with sparsity: the same word is rarely used. The most used word is in 22% of the comments. Result doesn't change if we increase the number of comments from 9308 to 46609.



| Comments | Unique words (dictionary length) | Descriptive Statistics: word used in how many comments? |
|---|---|---|
| 9308 | 11108<br>(6001 after filtering out words that occur less than in 2 comments)<br>See attached excel file: Dictionaries.xlsx - 10k sheet | mean 16.369553 (0.18%)<br>std 71.812963<br>min 1.000000<br>25% 1.000000<br>50% 2.000000<br>75% 6.000000<br>max 2075.000000 (22.29%) |
| 46609 | 23529<br>(12775 after filtering out words that occur less than in 2 comments)<br>See attached excel file: Dictionaries.xlsx – 50k sheet | mean 38.875473 (0.08%)<br>std 252.560218<br>min 1.000000<br>25% 1.000000<br>50% 2.000000<br>75% 7.000000<br>max 10429.000000 (22.38%) |

- The most frequent words should be considered as stop words since they are used as top 10 words across all topics. See Topics.xlsx
-

**TO DO:**

- ~~Change data source granularity adding parent and child comments together. Current analysis bases on randomly picked single comments independently from their position on the thread (comments created utc from 2016-07-01 to 2019-06-30). This add more noise to the messages already highly susceptible to intertwined conversations and incoherence.~~

already done or should be done.

- Improve code pipeline to allow for more flexibility (tuning params) and reusable code
- ~~Implement LDA genism multicore to speed up model training~~
- ~~Run topic modeling on training data, with all combination of alpha from [0.01, 0.05, 0.1, 0.5, 1, 5, 10, 50] and beta from [0.01, 0.05, 0.1, 0.5, 1, 5, 10, 50]~~
- ~~Find the sweet spot/the most informative words adjusting *filter_extremes*~~
- ~~Remove internet jargon/slang/acronym >> **Rebekah, you mention a package used in twitter**~~
- ~~Find informative comments based on comments length~~
- ~~Divide the data in 80% training and 20% test~~ What are we predicting?
-

**ALTERNATIVE PATHS:**

- Data source granularity for sense-making. More about it here: https://doi.org/10.25300/MISQ/2018/13239
- Validation using *MTurk* >> labelling based on % of number of people
- Co-occurrence patterns using word embeddings for unsupervised learning (heating patterns)
-

BR,

**Morena Rivato**
Research Assistant
MSc Business Intelligence

Mail: mor@mgmt.au.dk

**Dep. of Management**
Aarhus BSS, Aarhus University
Fuglesangs Allé 4, 2623-D203
DK - 8210 Aarhus V
http://www.mgmt.au.dk

MAPP CENTRE – RESEARCH ON VALUE CREATION
IN THE FOOD SECTOR
DEPARTMENT OF MANAGEMENT
AARHUS UNIVERSITY

**From:** Rebekah Brita Baglini <rbkh@cc.au.dk>
**Date:** Wednesday, 29 January 2020 at 16.47
**To:** Morena Rivato <mor@mgmt.au.dk>
**Cc:** Marco Hubert <mah@mgmt.au.dk>
**Subject:** Re: Meeting catch up

Hi Morena,

It was nice to meet with both of you yesterday! I will await access instructions from IT and hopefully have a chance to look over the code in the next couple weeks. Let's plan to follow up when you're back from vacation in mid-Feb.

Best,
Rebekah

**Rebekah Baglini**
Assistant Professor in
Linguistics

rbkh@cc.au.dk

**Interacting Minds Centre**
Aarhus University
Jens Chr. Skous Vej 4, 1483-323

8000 Aarhus C, Denmark

---

**From:** Morena Rivato <mor@mgmt.au.dk>
**Sent:** Wednesday, January 29, 2020 3:33 PM
**To:** Rebekah Brita Baglini <rbkh@cc.au.dk>
**Cc:** Marco Hubert <mah@mgmt.au.dk>
**Subject:** Meeting catch up

Dear Rebekah,

Thank you for the meeting yesterday.

For now, I only share with you yesterday's presentation where you can also find the links to the my github code (under the NLP pipeline headline) and to the data source.

I asked the IT to give you access to the server. Let's see how it goes.

I will be on vacation until February 9th with limited access to the internet.
We can get in contact after this date to see if this project could be something of your interest and I will be available to provide all the support needed.

Best Regards,

---

**Morena Rivato**
Research Assistant
MSc Business Intelligence

Mail: mor@mgmt.au.dk

**Dep. of Management**
Aarhus BSS, Aarhus University
Fuglesangs Allé 4, 2623-D203
DK - 8210 Aarhus V
http://www.mgmt.au.dk

AARHUS BSS

MAPP CENTRE – RESEARCH ON VALUE CREATION
IN THE FOOD SECTOR
DEPARTMENT OF MANAGEMENT
AARHUS UNIVERSITY

AACSB ACCREDITED AMBA ACCREDITED EFMD EQUIS ACCREDITED