This readme is about two .ipynb files (engagement.ipynb and social support.ipynb)

# Engagement analysis:

## Step 1: prepare data

We applied two data-frame to construct our data-set (‘df’) and corpus (‘corpus’):

data/Official\_Provincial\_Weibo\_From\_20191201\_To\_20200816.xlsx

data/0820.xlsx

## Step 2: model and result

Note: we apply likeP, commentP and forwardP as the labels rather than the raw like, comment, forward count. The reason is to normalize the data and scale it into [0, 100] range. It might be better to have an idea of “how likely the post will be liked/commented/forwarded” based on a value between 0 and 100. For example, a post with commentP value 10 can indicate the post is much less likely to be commented compared to a post with commenP value 90.

### Static like/comment/forward prediction

A standard way to do time series prediction is to try linear regression.

The feature here is tfidf.

Three different models has been tested: SGDRegressor (linear regression), RandomForestRegressor (random forest), MLPRegressor (neural network). Grid search has been done to find out the best hyperparameters. Results can be found as below:

|  |  |  |  |
| --- | --- | --- | --- |
| mean square error | like | comment | forward |
| SGDRegressor | 6.68 | 4.25 | 4.24 |
| RandomForestRegressor | 5.62 | 3.48 | 3.46 |
| MLPRegressor | 6.11 | 3.65 | 3.68 |

Notes: the table shows RandomForestRegressor is of the best performance. Meanwhile, comment and forward are identical across the models. The reason is they are almost of the same value in our dataset. (once it is a forward, it will be counted as a comments).

### Dynamic like/comment/forward prediction

A standard way to do time series prediction is to try ARIMA, AR, or MA. Here, we tried ARIMA and MLP(neural network) as our model candidates.

Note: Unlike previous item, we apply to the raw like, comment, forward count as the label. The reason is about the computation. (think about modeling the percentage value is too difficult to optimize in time series prediction). Meanwhile, the key factor in time series model is the value of lag. We tried [1, 3, 7, 14, 30] day then apply grid search to find out the optimal one. (kind of more complex in ARIMA, since it requires (p,d,q) as parameters)

|  |  |  |  |
| --- | --- | --- | --- |
| mean square error | like | comment | forward |
| ARIMA | 3.57 | 2.46 | 2.46 |
| MLP | 2.86 | 2.65 | 2.89 |

Notes: the table shows MLP of the best performance of like prediction whikle ARIMA performances better in comment & forward prediction. Meanwhile, comment and forward are identical across the models. The reason is they are almost of the same value in our dataset. (once it is a forward, it will be counted as a comments).

# Social support analysis:

## Step 1: prepare data

We applied one data-frame to construct our data-set (‘df’) and corpus (‘corpus’):

data/Official\_Provincial\_Weibo\_From\_20191201\_To\_20200816.xlsx

## Step 2: model and result

### Unsupervised topic modeling

A standard way to do LDA and NMF.

After set n\_topic = 10, we have the following topics:

**LDA result:**

Topic 0:(maybe “pandemic control”?)

疫情 防控 工作 肺炎 新型 新闻 新冠 发布会 冠状病毒 企业

Topic 1:(maybe “information”?)

视频 微博 中国 新闻 生活 近日 武汉 时间 人民 医院

Topic 2:(maybe “local daily event”?)

发布 今天 天气 工作 地区 全省 发展 全国 建设 我省

Topic 3:(maybe “pandemic statistics”?)

病例 确诊 新增 肺炎 累计 报告 出院 输入 境外 新型

Topic 4:(maybe “development impact”?)

链接 网页 发展 全国 企业 中国 工作 建设 今年 项目

**NMF result:**

Topic 0:

病例 确诊 新增 累计 报告 出院 输入 境外 肺炎 日时

Topic 1:

链接 网页 工作 发展 服务 建设 活动 项目 发布 我省

Topic 2:

候诊 医院 人市 人民 中医医院 发热 门诊 中心医院 第二 医科大学

Topic 3:

视频 微博 发布 今天 中国 全国 全省 地区 时间 天气

Topic 4:

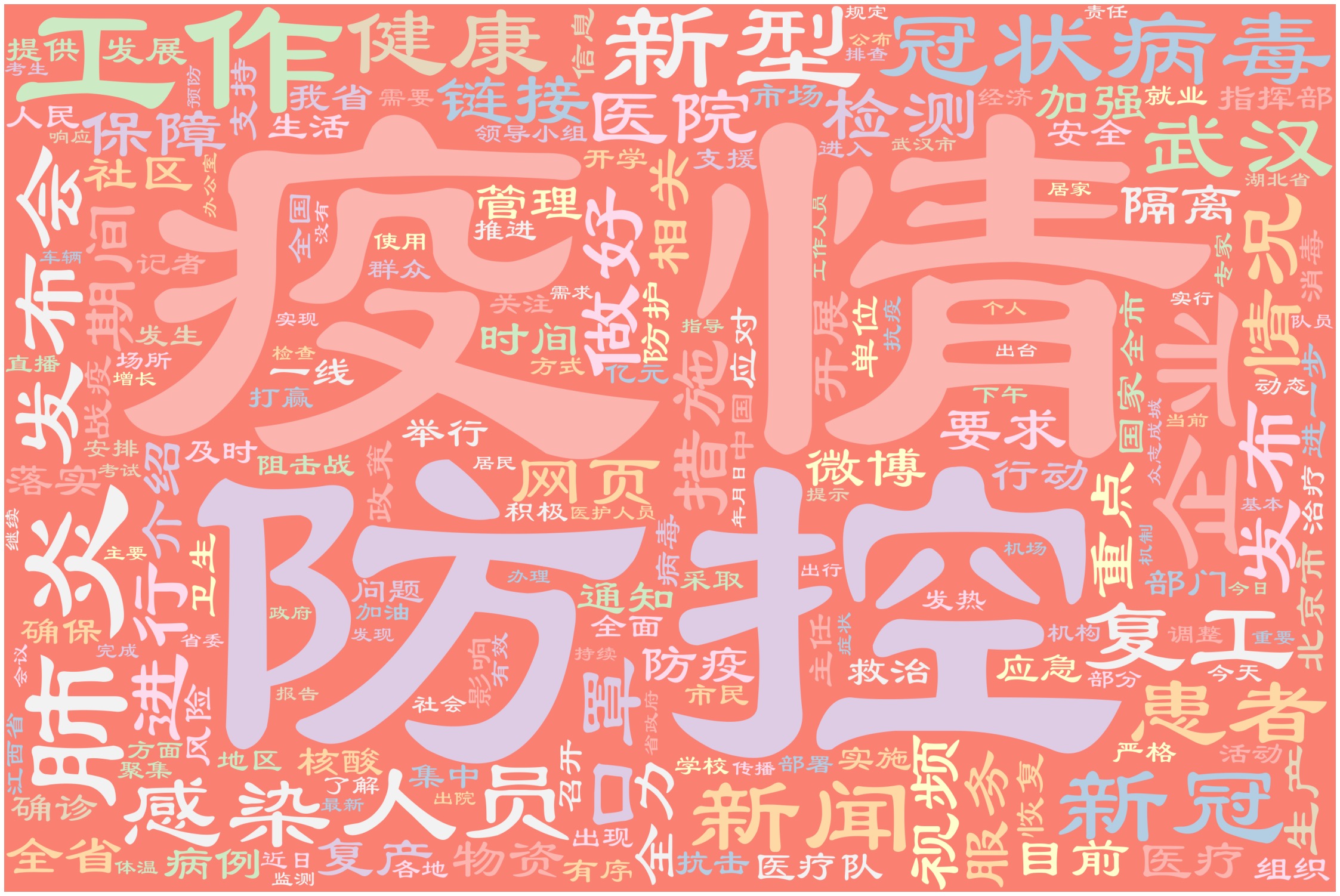
疫情 防控 肺炎 工作 企业 冠状病毒 新型 人员 患者 新冠

Note: I feel LDA and NMF results are very similar and I did find some reasonable result.

### Word cloud visualization of LDA result

Topic 0:(maybe “pandemic control”?)

疫情 防控 工作 肺炎 新型 新闻 新冠 发布会 冠状病毒 企业



Topic 1:(maybe “information”?)

视频 微博 中国 新闻 生活 近日 武汉 时间 人民 医院



Topic 2:(maybe “local daily event”?)

发布 今天 天气 工作 地区 全省 发展 全国 建设 我省



Topic 3:(maybe “pandemic statistics”?)

病例 确诊 新增 肺炎 累计 报告 出院 输入 境外 新型



Topic 4:(maybe “development impact”?)

链接 网页 发展 全国 企业 中国 工作 建设 今年 项目

