**Flu Spread @Korea Project Week 3, Team XLARGE**

**Obtained Data Source List**

1. Flu Data

This project requires data about actual occasions of flu. This data is mandatory to establish and validate our model.

The minimal temporal granularity of open data is a week and requires manual processing as they are form in documents. Even closed data is not accumulated in real-time. Therefore, it seems that no flu data is suitable for real-time prediction and validation. This is a crucial problem on progress of this project. It is important to decide whether this data is enough to continue this project.

We assume that data by time and data by location is dependently changes according to a flu spreading model. Therefore, data of fine spatial granularity alone and fine time granularity alone cannot be merged, as merging requires the assumption of independence.

* 1. Sample of Medical Data

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Site | Sample Cohort DB, National Health Insurance  nhiss.nhis.or.kr/bd/ab/bdaba002cv.do | | | Data For\* | | Flu, Cold |
| Gra. | Duration | ‘06~’16 |
| Status | Closed ($630) | Update | Annual | Temporal | Day? |
| Type | Remote Database | Domain | 1M Sample | Spatial | County? |

\* National Health Alarm Service uses this data to predict cold.

* 1. Student Illness Data

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Site | Student Health Info, Ministry of Education  schoolhealth.kr/shnhome/bbs/bbs01002v.php | | | Data For | | Flu |
| Gra. | Duration\* | ‘15~’17 |
| Status | Open | Update | Weekly | Temporal | Week |
| Type | Document, Num. | Domain | Students | Spatial | Province |

\* Data from vacation period, including Jan-Feb, is omitted.

* 1. Influenza Occurrence Rate by Weeks

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Site | Weekly Report on Infections, CDC Korea  cdc.go.kr/CDC/info/CdcKrInfo0402.jsp?menuIds=HOME001-MNU1132-MNU1138-MNU0045 | | | Data For\* | | Flu |
| Gra. | Duration | ‘15~’17 |
| Status | Open | Update | Weekly | Temporal | Week |
| Type | Document, Num. | Domain | Total | Spatial | Nation |

\* Aggregation by virus types is given.

* 1. Patient Sample

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Site | Patient Sample, Healthcare Bigdata Hub  opendata.hira.or.kr/op/opc/selectPatDataAplInfoView.do | | | Data For | | Flu, Cold |
| Gra. | Duration | ‘09~’16 |
| Status | Closed ($270\*) | Update | Annual | Temporal | Day? |
| Type | Dataset | Domain | 1.4M Sample | Spatial | County? |

\* The price is per data of 1 year.

* 1. Disease by Province

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Site | Disease by Province, Healthcare Bigdata Hub  opendata.hira.or.kr/op/opc/selectOpnsApi.do?sno=711 | | | Data For | | Flu, Cold |
| Gra. | Duration | ‘??~’16 |
| Status | Use by Permission | Update | Annual | Temporal | Year |
| Type | API | Domain | Total | Spatial | Province |

1. Social Media Data

We have searched flu posts, but there are less than 50 posts a year on social media in Korea. This is because in Korea, when some people got flu usually they say “I have a cold”. Therefore, we tried to find “cold” from SNS data. Sure there were more posts compare to flu, but the problem is there is no suitable actual data about cold (see Section 1) to correlate with.

* 1. Twitter Search/API

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Site | Twitter Search / Twitter API(address below)  <https://developer.twitter.com/en/docs/tweets/search/overview> | | | Data For | | Cold |
| Gra. | Duration | ‘06~’17/  1 week |
| Status | Open/Closed | Update | Real time | Temporal | Day |
| Type | Webpage/API | Domain | Posts | Spatial\* | National |

\* The posts are not enough if we add location information to the search criteria.

* 1. Instagram API

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Site | Instagram Search API  https://www.instagram.com/developer/ | | | Data For | | Cold |
| Gra. | Duration | ?? |
| Status\* | Closed | Update | Real time | Temporal | Real time |
| Type | API | Domain | Posts | Spatial | 1~5km |

\* Access Token is given after submission of apps running by companies.

* 1. Facebook API

Facebook removed search-by-post functionality on their API version 2.

1. Search Data

We found some data that mean statistics based on searching words. We can use search words ‘독감(Flu)’, ‘감기(cold)’, etc.

* 1. Google Trends

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Site | Statistics about searching words, Google Trends  https://trends.google.com/ | | | Data For | | Flu, cold |
| Gra. | Duration | ‘13~’17 |
| Status | Open | Update | Weekly | Temporal | Week |
| Type | Graph, Data set | Domain\* | Search Count | Spatial | Province |

\* Relative Value

* 1. Naver KnowledgeIN Search Data

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Site | KnowledgeIN (a Q&A site) Search Count, Naver  kin.naver.com | | | Data For | | Flu, Cold |
| Gra. | Duration | +15 yrs |
| Status | Open | Update | Realtime | Temporal | Day |
| Type | Webpage | Domain | Q&A Count | Spatial | National |

1. Weather Data

Weather data is the least important data among the three categories, as it is not directly related to flu patients. On the last week, we successfully demonstrated acquisition of interpolated weather data of arbitrary coordinates in Korea from KMA Automatic Weather Stations(AWS). We can also collect per-station data if needed, whether real-time or previous, as they are all open.