2022 Data Science Competition

Kathy Pai, Sungkeun Kim

Summary

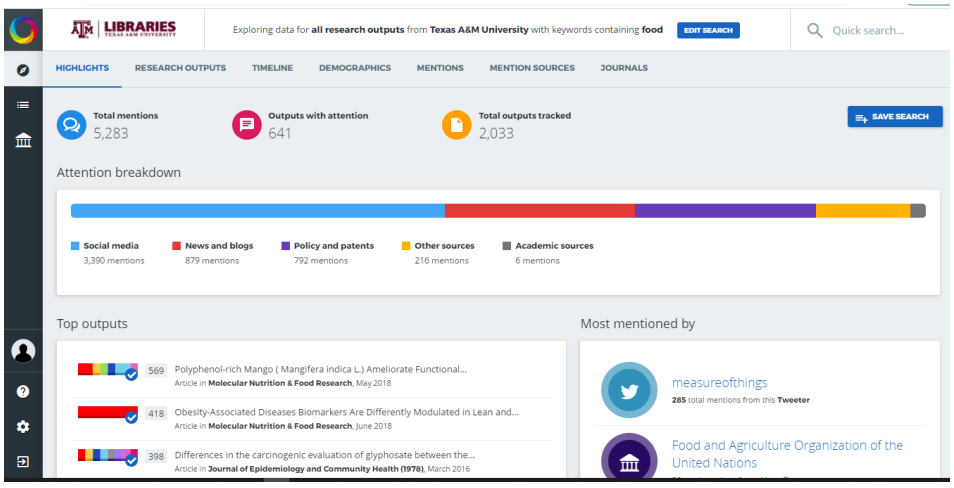
As computation power grows in recent decades, Data science and its related theories such as Artificial intelligence and Machine learning are revived and actively being researched. In this competition, we try to find interesting questions or applications and solve them with the provided source of data.

Problem Statement

The goal of this competition is to describe and visualize the network of TAMU's research and show patterns of research. In other words, we want to show what areas of research that our school has a great impact on society and what areas could be strengthened to address social challenges. We believe that our work shows insight or recommendation of future success of research at our school. The target audience of this project will be university leaders or the public.

Dataset

The data we're using is downloaded from the Altmetric website, and it's about all the research in TAMU in all the departments. It includes the research publication title, the departments that worked on the publication, the collaborators of these works, the funding source of the publication, and whether it's mentioned in any social media. Overall, we want to preprocess the title of the publication and use multilabel classification to train the title with departments, collaborators, and funders. We use 1000 dataset for testing (only using title of publication and department of publication to train) since the whole dataset is large.



<Figure 1. Almetric website>

Methodology

For our project, we assume the user provides a short description of his/her interest in a specific topic, which will be used as input to the machine learning model, and it'll provide some departments, collaborators, and funders related to this subject as output. We also create a visualization graph using the data. This will give users a better understanding of what departments are likely working on research in this area, what other schools are also working in this area, and where people could possibly get funding for this type of research.

Modeling and Analysis

For now, we have some code that cleans up the title to remove unnecessary words then vectorizes it to use as input, and we're experimenting with different classifiers now. There are a few things that need to be done. First, we want to find whether there is a better way to preprocess the title text input. Second, we test different classifiers and generate graphs based on accuracy, and we can use the classifier with the highest accuracy score to train our data and produce the output mentioned earlier in this context. Lastly, we need to look for techniques/tools to visualize our data since it's an essential output for this competition. The current challenge is understanding how to input data into these classifiers with adequate parameters, finding the classifier suitable for our needs, and producing visualizations of these data for the competition.
First, we downloaded the data from the Almetric website and preprocessed as shown in Figure 2.

	B	C	D	E	F	G	H	I	J	K	L	M	N
1	Departments	Affiliations (GRIC	Funder	News mentions	Blog mentions	Policy mentions	Patent mentions	Twitter mentions	Facebook mentic	Wikipedia mentic	Google+ mentior	Reddit mentions	Video mentions
2	Atmospheric Scie	California Institut	Welch Foundatio	719	47	3	1	37517	32	6	0	22	6
3	Texas A&M; Med	Aga Khan Univer	Canadian Institut	315	35	3	0	12140	520	9	7	20	30
4	Atmospheric Sciences; College of Geosciences; Tex			95	17	0	0	9553	46	8	2	17	3
5	College of Agricu	Chosun Universi	National Institute	510	47	0	0	3210	53	6	0	2	14
6	Texas A&M; Computer Science an		Laura and John A	364	238	6	0	2401	60	27	25	10	2
7	Cardiovascular Disease: A Presidential Advisory Fro			470	29	1	0	1045	112	23	8	3	39
8	College of Medic	Guy's Hospital; I	Australian Resea	178	10	0	0	3624	3	11	0	2	0
9	College of Libera	Baylor University	Arts and Humani	143	34	0	0	2952	20	14	0	6	0
10	Aerospace Engineering; College of Engineering; Tex			356	25	0	0	296	1	4	0	2	1
11	College of Scien	Autonomous Uni	Deutsche Forsch	154	36	13	5	3043	2	22	0	1	3
12	Atmospheric Scie	Lawrence Liverm	Directorate for G	242	23	1	0	704	4	0	0	4	0
13	f tetrachlorvinpho	International Age	National Cancer	271	70	6	0	1218	181	6	47	5	2
14	School of Public	Karolinska Institu	AFA Insurance (S	234	17	0	0	709	16	1	2	0	0
15	ascular Nutrition	Cleveland Clinic;	Florida Internatio	210	8	0	0	761	114	3	6	4	2
16	Physics and Astr	Ames Research	Directorate for M	229	17	0	0	265	13	18	62	0	1
17	School of Public	University of Pennsylvania; Utah V		91	18	0	0	1613	22	0	6	13	2
18	Physics and Astr	Canadian Institut	Directorate for Ec	197	18	0	0	126	2	6	0	2	3
19	School of Public	Hospital for Sick	Canadian Institut	146	35	1	1	754	142	4	10	2	38
20	College of Scien	Autonomous University of Madrid;		108	19	5	0	1473	8	31	0	0	1
21	Texas A&M; Physics and Astronon	Economic and Sc		160	24	0	0	875	22	35	3	0	2
22	Texas A&M; Colle	Carnegie Mellon University; ETH Z		178	10	1	0	323	4	0	0	0	0
23	Civil Engineering	Delft University o	European Comm	190	2	0	0	38	0	0	0	0	0
24	Atmospheric Scie	Texas A&M University; University o		103	22	5	0	1151	9	1	1	0	0

<Figure 2. Downloaded and preprocessed dataset>

Then, we use TfidfVectorizer and CountVectorizer to transforms input as shown in Figure 3 and Figure 4.

```
BinaryRelevance - accuracy score: 0.18090452261306533, hamming loss: 0.08384990646663977

ClassifierChain - accuracy score: 0.18592964824120603, hamming loss: 0.02512562814070352

LabelPowerSet - accuracy score: 0.2663316582914573, hamming loss: 0.026372739610461063

KNeighborsClassifier - accuracy score: 0.17587939698492464, hamming loss: 0.01819315555881598

DecisionTreeClassifier - accuracy score: 0.24120603015075376, hamming loss: 0.024758830649598357

RandomForestClassifier - accuracy score: 0.19597989949748743, hamming loss: 0.017532920074826687
```

<Figure 3. Result of testing with different multilabel classification method and used TfidfVectorizer to transform input>

```
BinaryRelevance - accuracy score: 0.21608040201005024, hamming loss: 0.03297509445035396

ClassifierChain - accuracy score: 0.22613065326633167, hamming loss: 0.01797307706415288

LabelPowerSet - accuracy score: 0.3417085427135678, hamming loss: 0.021604372226093974

KNeighborsClassifier - accuracy score: 0.12562814070351758, hamming loss: 0.018303194806147526

DecisionTreeClassifier - accuracy score: 0.27638190954773867, hamming loss: 0.020724058247441588

RandomForestClassifier - accuracy score: 0.16080402010050251, hamming loss: 0.01694604408905843
```

<Figure 4. Different multilabel classification method after preprocessing the input to remove some possibly unnecessary words and used CountVectorizer to transform input>

Virtualization and Interpretation

Using the above classification, finally we produce the visualized result that shows the most relevant research topic for the given input. We can see “Health Science Center” is the most active and makes a lot of publications with fundings in Texas A&M University.

<Figure 5. Visualization of the classification>

Conclusion

Data science, AI, and machine learning is the one of the dominant research areas that help our real life impactfully and it is possible with help of on-going researches and competitions such as the 2022 Data Science Competition in TAMIS. In this competition, we try to find interesting questions or applications and solve them with the provided source of data. We could visualize the network of TAMU's research and show patterns of research. We showed what areas of research that our school has a great impact on society and what areas could be strengthened to address social challenges. From our experiment and analysis, our work recommends pursuing study about Health science with the Health Science center.

Links to Executable code and supplementary materials

We stored executable code and supplementary materials in the git repository in the link below.
https://github.com/ksungkeun84/TAMIDS