Final Project Report

Zhewei Wang

1 Data

Firstly list the data we have and the useful information in them.
The data files we have are:
1. program.csv, which contain program ID, program objectives, program start date and end date,
2. project.csv, which contain project ID, corresponding organization ID, corresponding program ID, status, project start date and end date, project objectives.
3. contractor.csv, which contain project ID and the corresponding organization ID.
4. publications.csv, which contain publication ID and the corresponding project ID, publication description, acceptance date.
5. organization.csv, which contain organization ID.
6. report.csv, which contain report ID and corresponding project ID, report details.
7. publication_author.csv, which contain publication ID and corresponding person ID.
8. person.csv, which contain person ID.

2 Problem Definition

At the beginning our idea is for a given project, find out the appliers(person), and then find out the projects the appliers completed and the papers the appliers published. Based on these information train a machine learning model to predict the given project will be successful or not. However we don't have the information that the projects are successful or not. Then we simplify our case as to predict how many papers will be published on this project, which is an important criterion for project evaluation.
According to project.csv, the information we can get is the organization corresponding to the project other than persons. We can find out the person

who did this project by finding out the papers published on this project and use the papers' authors as the apply persons. Our plan is to finish a simple version first. So we simply use organization as the applier. The model to us now is: given a project and its apply organization, we find out the projects that the organization finished and the papers the organization published, and then give our prediction by a machine learning model.

3 Implement

Firstly we find out completed projects, which are 8388 in our data. And then we need to find out the projects whose apply organizations have projects finished and papers published before, otherwise it is useless for our predict is given based on the past publication record. Now we have 1993 project satisfy these constraints, which are our whole data.
Now we should design features. The features we used are the documents similarity between the objectives of given project and the description of papers published by the apply organization, and the publication quantity.
The method we used to measure the documents similarity is Word Mover's Distance*. For two given documents, this method will remove the stop words from the two documents first, then find the vectors of the rest words in the pretrained word embedding matrix. The distance of these two documents is the distance that the words of one document move to reach the words of the other document. For The whole implement we followed *http://vene.ro/blog/word-movers-distance-in-python.html*.
In our case, the word embedding matrix we used is Google News Vectors. when documents distance are calculated some words are not exist in the matrix. One situation is special numbers. In Google News Vectors, only the common numbers such like '1', '5' have word embedding vectors. For some numbers like '32' the word embedding is not exist. So we just ignore them. Another situation is some special words. At beginning we thought they were misspelled. Then we used spell checker which follow the implement of *http://norvig.com/spell-correct.html*. After the word spell check some of the words are still misspelled. By checking these word on the internet, we found that these words are special words like 'eurostate' which is an organization, or the plural form such like 'analyses'. So we also ignore them.

After the ignorance, some documents have no words rest. Then the Word

*Kusner, Matt, et al. "From Word Embeddings To Document Distances." Proceedings of the 32nd International Conference on Machine Learning (ICML-15). 2015.

Mover's Distance method will give us 'nan'. We transfer the 'nan' to value 0.
Besides these two features we can add more features. Here we just keep it simple and finish the whole framwork first.
Then we keep two data structures. One is for the organization, as shown in Figure 1. And one is for whole 1993 projects.

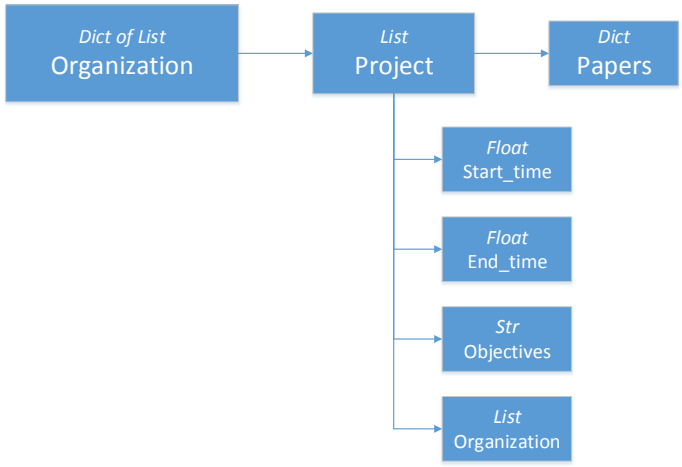

Figure 1: Data structure for organization.

The learning model we used is SVM. We use the python package scikit-learn. We separate our whole 1993 project as training data, validation and test data. And the ratio between the three data is $8:1:1$. After training, the SVM model give us the result of all 0. If we output the publication of these 1993 projects we can see as show in Fighure 3. Most the projects has no publications. It makes sense for SVM give us the prediction that are all 0. But it is not so useful when we consider the fact is most projects' publication is 0. Then we shrink our data to the projects whose appliers have papers published before and actually itself also have paper published. Now we only have 269 projects. Repeat our experiment on the new data, the result we got is shown Figure 4. The accurate is about 50%. It's not good

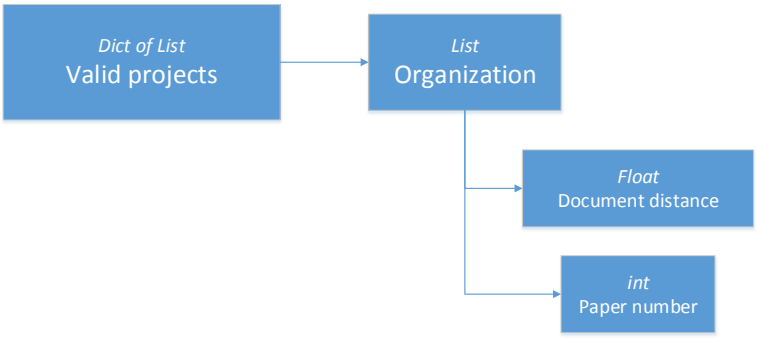

Figure 2: Data structure for whole 1993 projects.

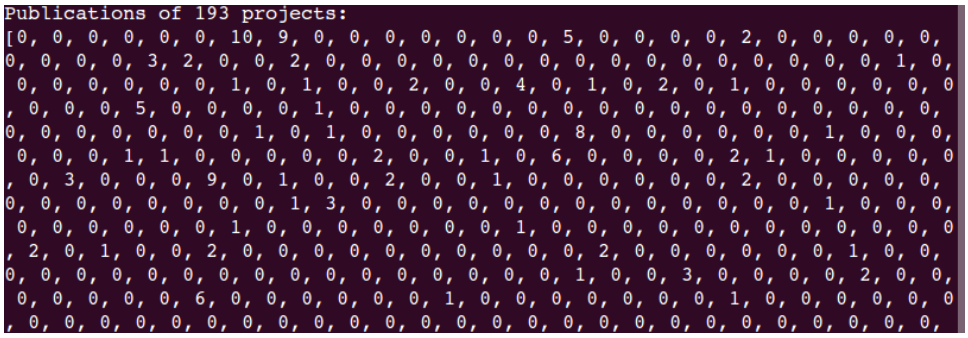

Figure 3: 1993 projects' publication whose appliers have paper published before.

```
Presict results are:
[ 1.  1.  1.  1.  1.  1.  1.  1.  1.  1.  1.  1.  1.  1.  1.  1.  1.  1.
  1.  1.  1.  1.  1.  1.  1.  1.  1.  1.]
Ground truth:
[  1.   1.   1.   3.   2.   1.  10.   1.   2.   1.   1.   1.   1.   1.   8.
   2.   2.   1.   1.   2.   1.   1.   1.   1.   1.   5.   1.   1.]
```

Figure 4: Prediction on 269 projects dataset whose appliers have paper published and itself also has publications.

especially when we consider that this is an unbalanced problem, publishing 1 paper is about 50% in these projects.

4 Improvement

The first question we should answer is whether the publication number is an important criterion to evaluate the project. Because we don't have the information that the project is evaluated as successful or not, we assume the more papers published, the better. However, we can see we have 8388 completed projects, and among them only 1267 projects have paper published, which are much fewer.

```
[7, 17, 10, 8, 4, 10, 14, 2, 14, 21, 2, 1, 3, 7, 10, 6, 21, 1, 6, 1, 1, 6, 16, 6
, 35, 8, 6, 2, 53, 2, 1, 6, 2, 1, 1, 3, 1, 53, 11, 1, 2, 9, 2, 14, 14, 1, 1, 65,
 8, 1, 2, 2, 5, 1, 20, 10, 22, 26, 4, 5, 21, 19, 24, 1, 29, 12, 5, 55, 1, 1, 25,
 2, 8, 2, 45, 18, 2, 11, 20, 1, 2, 82, 1, 38, 1, 1, 43, 1, 6, 61, 12, 10, 1, 18,
 5, 4, 2, 2, 8, 2, 3, 7, 2, 1, 1, 5, 1, 1, 8, 3, 2, 9, 7, 3, 1, 2, 46, 7, 10, 1,
 7, 2, 23, 3, 6, 24, 14, 10, 8, 8, 22, 4, 1, 3, 20, 6, 1, 6, 11, 2, 6, 1, 7, 1,
52, 19, 9, 1, 4, 6, 1, 4, 7, 1, 26, 5, 1, 6, 1, 1, 33, 2, 1, 6, 2, 1, 22, 21, 8,
 4, 2, 1, 1, 7, 89, 2, 10, 8, 28, 1, 8, 7, 1, 21, 1, 13, 2, 18, 1, 2, 17, 1, 6,
12, 2, 60, 2, 5, 1, 24, 75, 75, 6, 3, 38, 1, 3, 43, 16, 67, 1, 5, 2, 23, 6, 5, 6
, 2, 7, 4, 5, 16, 3, 152, 22, 1, 15, 6, 1, 1, 10, 3, 3, 63, 1, 5, 68, 12, 9, 20,
 4, 5, 1, 2, 1, 9, 9, 7, 8, 2, 1, 1, 1, 10, 8, 23, 12, 4, 1, 2, 7, 6, 1, 5, 1, 6
, 1, 7, 1, 1, 32, 11, 10, 7, 1, 9, 2, 6, 3, 2, 2, 7, 1, 2, 12, 8, 11, 5, 21, 8,
8, 2, 2, 6, 1, 1, 1, 9, 4, 15, 8, 2, 3, 21, 2, 14, 12, 36, 63, 1, 10, 8, 1, 7, 1
, 5, 7, 2, 17, 3, 26, 27, 2, 3, 1, 6, 10, 5, 30, 15, 8, 13, 51, 13, 32, 1, 1, 1,
```

Figure 5: Publications of the organizations who applied one of 1993 projects.

Our assumption is good appliers have a good chance will be successful in the next project, or good appliers with most publications will be most likely to publish papers in the next project. In our case it doesn't happen. Someone can argue that that's because the information we used to give prediction is not enough. But we always thought document similarity and publication numbers are two majority features. These are not like innovation or paper annual citation, which are used to give more detail information about pub-

lication quantity. These two major features themselves should give us good enough perditions. But we can see that most organizations have publications before from Figure 5. Then compare Figure 5. with Figure 3. we can see there is no direct connection between the publication in the past and the publication in the future because no wonder how many papers published before, there still no paper for the current project with large possibility. Consider the performance of SVM model, it is more likly only give us the prediction with the largest probability in the current situation. For example in Figure 3. the SVM's prediction is always 0 and in Figure 4. the prediction is always 1. It is just like $p(A|B) = p(A)$.

One explanation about why publication in current or future project (A) is irrelevant with the publication in the past (B) is the base of our assumption that publication is an important criterion for evaluation. If the appliers don't have to use their publication to earn good judgment, then no wonder 'good' appliers or 'bad' appliers, they don't pay attention on publications.

So I strongly doubt if publication quantity is used for evaluation, or as a major criterion used in evaluation. Only a positive answer be given for this problem, we can do the rest to improve our model.

The first thing we can do to improve our model is refine the appliers to person other than simply consider about organization. In reality EU knows who applied which project. To mimic people's process maximumly, we use the authors who published papers on this target project as the persons who applied this project. The results should be more accurate. And also, when we refine the appliers to person, the publications are also refine to authors other than organization.

Another feature we want to add is innovation. Consider if we have several proposals for a project, the proposal with more innovation should be more like to win and successful in the future. In our case we only have one proposal for each project. The way we try to use innovation is based on a assumption that in a program, the project is more innovate, more papers will be published. To check if this assumption is correct, we calculate the distance between a project and the rest projects in a program. And the result is shown as Figure 6. We can see it doesn't follow our expectation. We still need to find a way to bring in innovation.

Other features we are considering include citation of papers –which reflect the quality of papers, project similarity, etc.

I also attached some slides as appendix, which is my research log in this project and contain more details about the whole program framework.

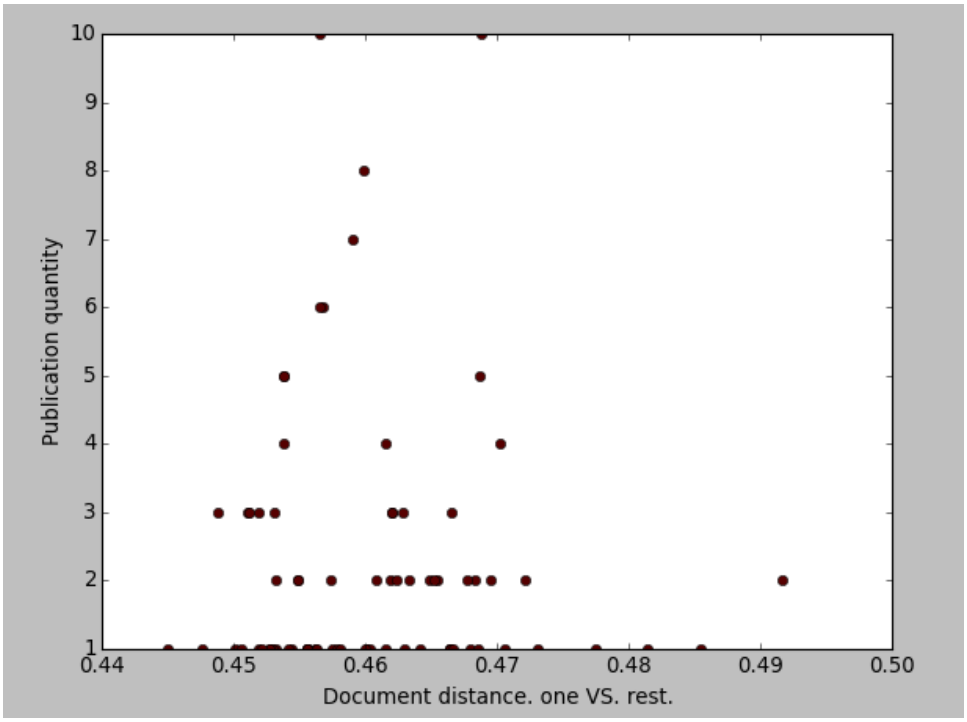

Figure 6: Publications quantity and innovation.

NLP Final Project

1. Preparing

1. Word2vec. ← Done
    a. http://vene.ro/blog/word-movers-distance-in-python.html
    b. Read GoogleNews-vectors-negative300 in memory
    c. Word embedding in embed.dat by replace ‘syn0norm’ with ‘syn0’.
    d. Generate a list in embed.vocab, then generate the index in vocab_dict
2. Test. ← Done
    a. d1 = "Obama speaks to the media in Illinois"
    b. d2 = "The President addresses the press in Chicago"
    c. Generate feature of these two documents.
    d. Pyemd to calculate d(doc_1, doc_2) = 0.63

2. Clean data.

Manually generate training data and test data?(No way.. Try automatically)

.csv files have some problem. Need to re-save to new .csv files.

1. Features of a project: choose a project, find organization, then find projects they did, papers they published…
    a. From publications by tracking who published, we can figure out who belongs to this organization. And every time when a project given, find all the papers published by the persons of this apply organization. Maybe it's better. But now let's keep it simple. Only use organization publication information.
    b. Actualy EU knows who did this project. To our case, we can use the authors who published in this project as the people who applied this project. By doing this we can replace apply organization with apply people. It should be more accurate. ← will do this later

C. For publications, didn’t consider about accepted time. Because they published on the corresponding project, and the status of the projects is completed, so they should be before the end_date of project. ← maybe not so correct

D. For paper dictionary, only keep the description in values.

2. Save data in data structures. Two data structures← Done

   a. Organization → project
   b. Project → papers & objectives & start_time & end_time

3. Completed project: 8388

3. Training and validation and test data

1. Generate data
    a. Find out the valid sample, aka, the projects that their appliers has papers before.
        i. 1993 valid samples
    b. A data structure to store the information.
        i. Thinking use the data structure of raw or an independent structure.
        ii. Choose the second choice… ← Done
    c. Features…
        i. Find out papers earlier than the start time.
        ii. Feature 1: distance between the papers and project. ← integrate word mover distance to current work. ← Done
            1. Some words are not exist in vocabulary because of misspelling.
                a. Spell checker
                b. After spell checking, ignore the incorrect words.
                c. Just found they are not misspell, they are some special words, like eurostat, or some plural form like analyses. So will just ignore them.
            2. Specific numbers like ‘32’ are also not exist in vocabulary.
                a. Simply ignore them.

4. Calculate Distance.

1. Sometimes it is nan. ←
    a. After removing special words, there is no words exist in one of the two documents. So just consider the similarity as 0.
    b.
2. All value between 0.45-0.62.

3. Distance, not similarity. So large number means not similar. But for nan we set it as 0, and for the purpose that set the features with the same length, we also fill the list with 0. The reason is 0 won't affect the learning model for weight*0 is still 0.
4. Forget to save the papers number published in the target project… ← go back to save it in pickle file.

5. Learning model

1. We have 1993 samples. In these samples only 269 has paper published.
   a. Too sparse.
   b. Meaningless if we do on 1993 samples…..

```
[ 0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  1.  7.  0.  0.
  0.  0.  5.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.
  0.  0.  0.  0.  0.  0.  2.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.
  4.  0.  0.  0.  0.  0.  0.  0.  0.  1.  0.  1.  0.  0.  0.  0.  0.  0.
  0.  2.  0.  1.  0.  0.  0.  1.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.
  0.  1.  0.  0.  0.  0.  1.  3.  0.  0.  0.  1.  0.  1.  0.  0.  0.  0.
  0.  0.  0.  0.  0.  0.  0.  0.  0.  1.  0.  0.  0.  0.  0.  0.  0.  2.
  0.  0.  0.  0.  1.  0.  0.  0.  0.  3.  0.  0.  0.  0.  0.  0.  0.  0.
  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.
  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  3.  1.  0.
  0.  0.  0.  0.  0.  0.  0.  1.  7.  0.  0.  0.  0.  0.  0.  0.  0.  0.
  0.  0.]
```

C. what if… only test the project which has paper published.

```
Presict results are:
[ 1.  1.  1.  1.  1.  1.  1.  1.  1.  1.  1.  1.  1.  1.  1.  1.  1.  1.
  1.  1.  1.  1.  1.  1.  1.  1.  1.  1.]
Ground truth:
[  1.   8.   1.   2.   5.   1.  10.   1.   1.   5.   1.   2.   1.   4.  10.
   2.   1.   3.   1.   1.   1.   4.   1.   3.   1.   1.   4.   1.]
```

improve:

1. Other learning model, such as RNN. ← :(
2. Predict how many papers each organization will publish on this project other than consider total of them. ← But we only know the total number. If we want the prediction on each applier, we should do this: project → publication → author → organization, then we know which publication belongs to which organization, we know in the total publications which ones are published by a specific organization.
3.

1. Innovative. The more innovative, the better. But innovative projects have more paper published?... ← check this first, then consider if use it as a feature.
    a. projects that are finished & projects that have paper published <-- 1267
    b. Distance between the project and the rest projects
    c. regression

We also have some simplification at here. We assume the paper and the corresponding project are very similar. So we didn’t compare the target project with past project, just with past papers.

Tips:

1. Data structure is important.
2. Project to org is one to multiple, org to project is also one to multiple. So need to use list().