Quick Overview

Goal: Study the evolution of topics funded by NSF in the field of AI.

Keywords: "artificial intelligence" OR "machine learning" OR "neural network" OR "computer vision" OR "speech recognition" OR "natural language processing" OR "deep learning"

Dataset: 10 years of NSF grant data (2010-2020) obtained with the above keyword search on the NSF website. Around 6626 abstracts.

Technique/Methodology:

1. Latent Dirichlet Allocation (LDA) for modeling topics based on grant abstracts.
2. Word embeddings to search for related words

Examples of Topic Model Visualizations

You can view the topic model visualizations for each of the 10 years by downloading the html files located here. Once downloaded, you can open these files in your browser.

- The circles labeled with numbers are topic clusters from the model in decreasing order of dominance. Where 1 labels the most dominant topic.

- The slider on the top right changes the “lambda” value. When lambda=1, it shows the most common words in the cluster (when you click on a cluster), as lambda is decreased, it shows words more specific (relevant) to that cluster

Word Embeddings

Word embeddings can be learned from data in order to look for related terms. The pictures below show words closest to “machine learning”, “deep learning” and “artificial intelligence” respectively in the dataset.

```
model.wv.most_similar("machinelearning")
```

```
[('deeplearning', 0.8922557830810547),
 ('physicsbased', 0.8208704590797424),
 ('suite', 0.8199580311775208),
 ('iterative', 0.8193440437316895),
 ('analytical', 0.8190265893936157),
 ('communicationavoiding', 0.8049052357673645),
 ('analytic', 0.8045275807380676),
 ('secondarily', 0.8001391887664795),
 ('employed', 0.7993574738502502),
 ('dataanalytics', 0.7974909543991089)]
```

```
model.wv.most_similar("deeplearning")
```

```
[('ensemble', 0.9145339727401733),
 ('modelbased', 0.9006441831588745),
 ('machinelearning', 0.8922556638717651),
 ('generic', 0.8784470558166504),
 ('graphbased', 0.8776005506515503),
 ('lossless', 0.8756814002990723),
 ('probabilistic', 0.8700089454650879),
 ('compression', 0.8625925183296204),
 ('parametric', 0.860786497592926),
 ('multiresolution', 0.8595075607299805)]
```

```
model.wv.most_similar("artificialintelligence")
```

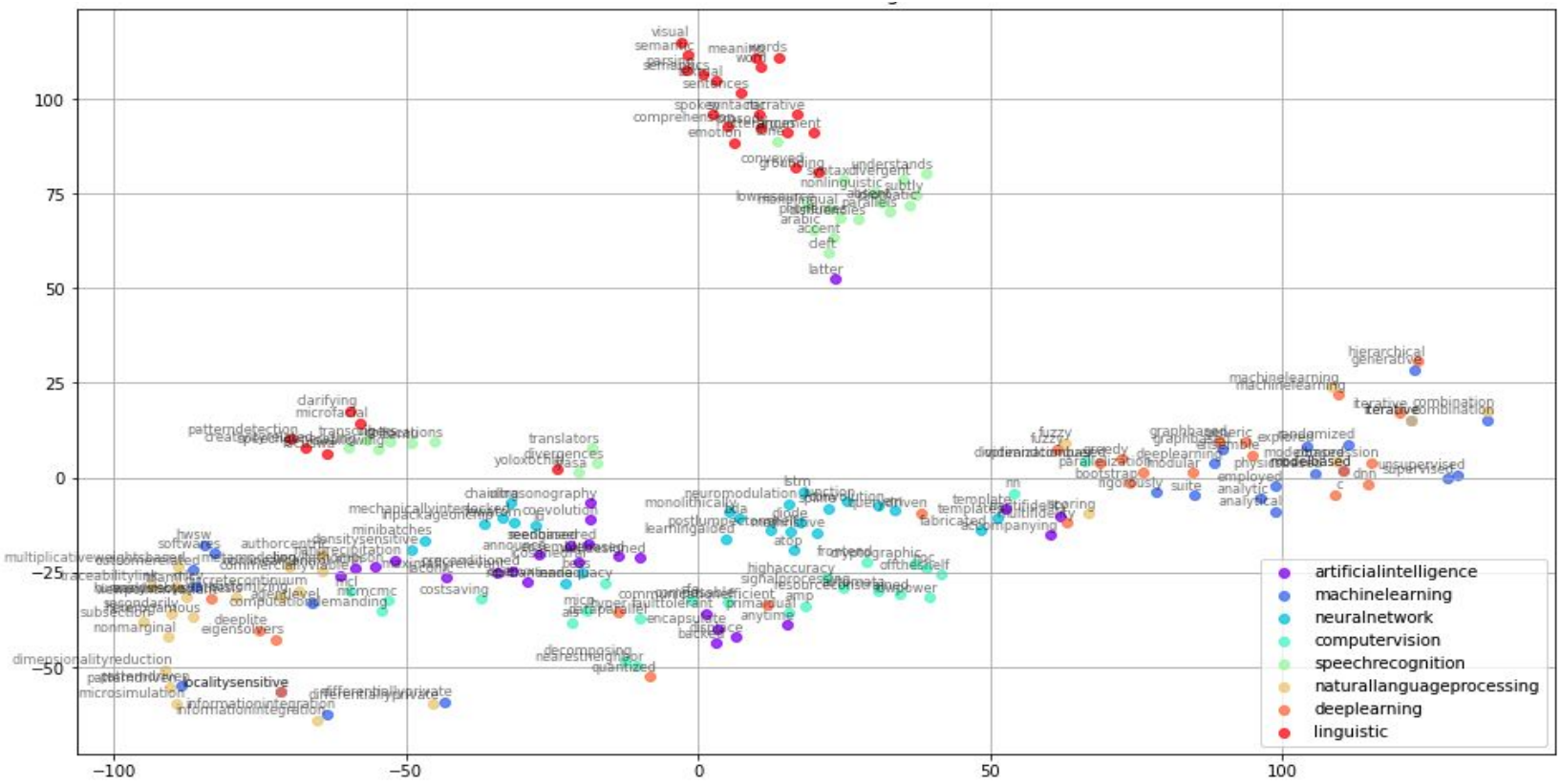
```
[('machinelearningbased', 0.7950197458267212),
 ('gamification', 0.7941629886627197),
 ('conceptualize', 0.786695659160614),
 ('devised', 0.7715743780136108),
 ('qa', 0.7684727311134338),
 ('repulsion', 0.7643722295761108),
 ('differentiallyprivate', 0.7621873617172241),
 ('expressive', 0.7616567015647888),
 ('naïve', 0.7581421136856079),
 ('braincomputer', 0.7533937692642212)]
```

Insights by Directorate

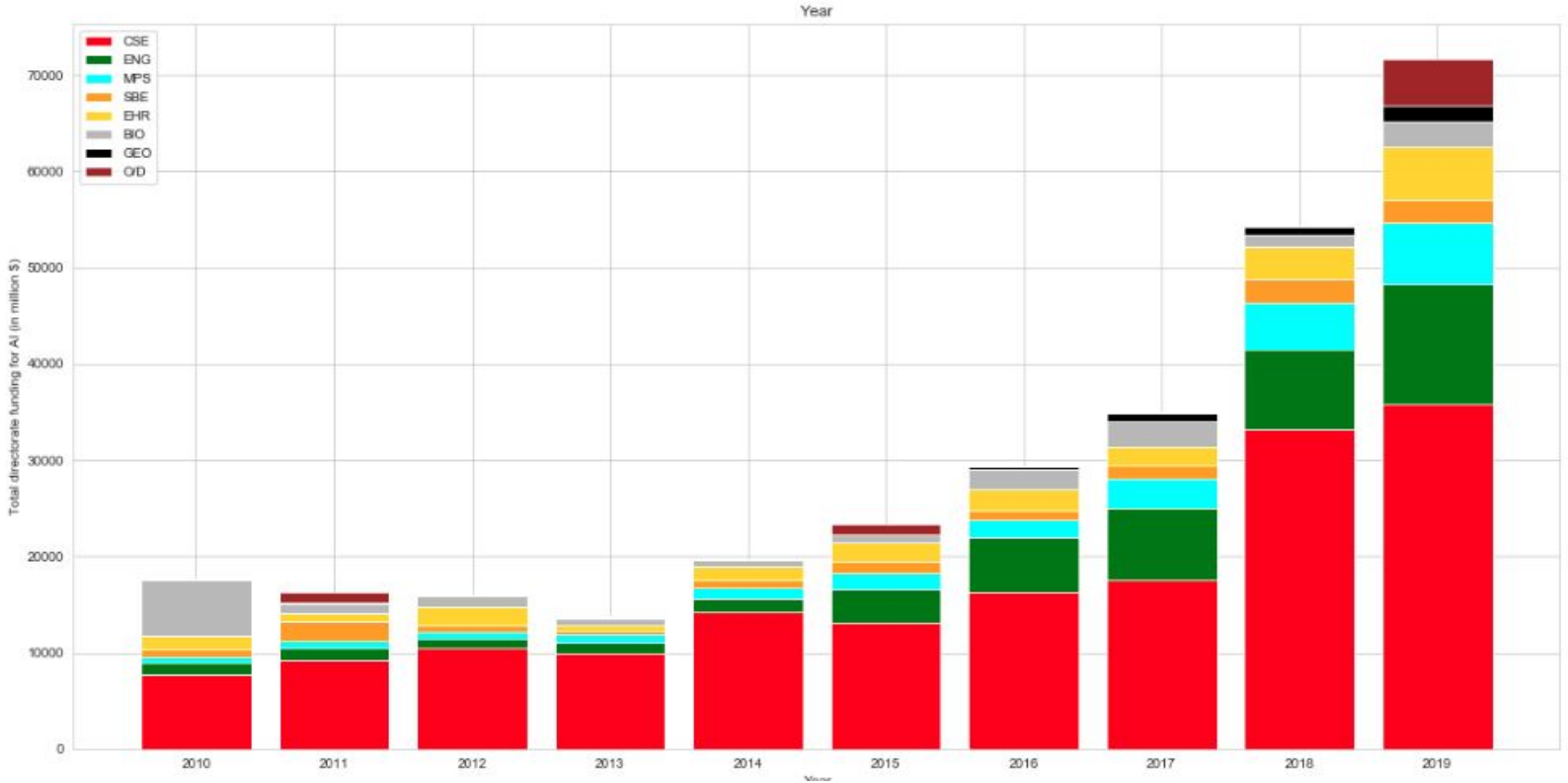
I used directorates as a placeholder for the type of “industry”. NSF has 8 main directorates that fund research in AI-

1. Biological Sciences (BIO)
2. Computer Science and Engineering (CSE)
3. Engineering (ENG)
4. Geosciences (GEO)
5. Mathematical and Physical Sciences (MPS)
6. Social, Behavioral and economic Sciences (SBE)
7. Education and Human Resources (EHR)
8. Office of the Director (OD)

Clustering of similar words



Funding per year by directorate



Number of awards per year by directorate