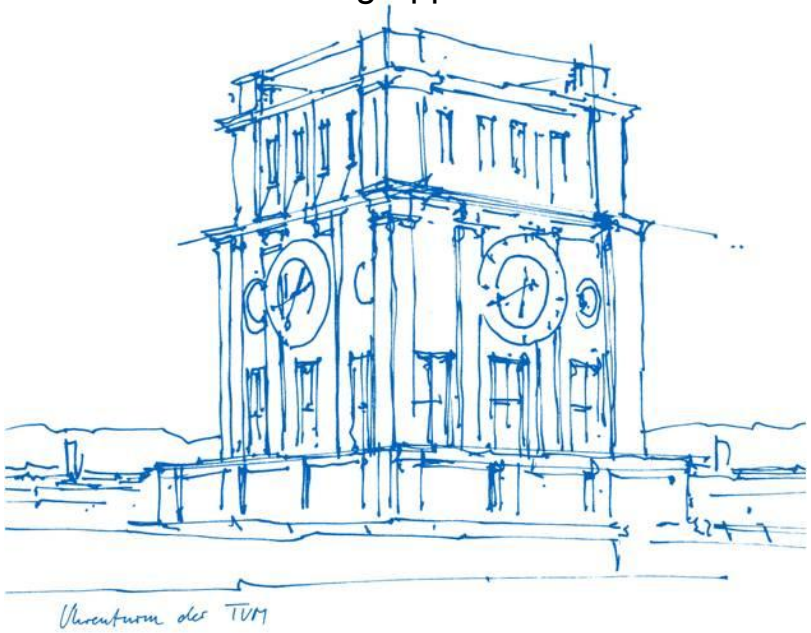
Opinion Mining Lab Group 1.3

Topic: Weakly Supervised Aspect Extraction Using a Student-Teacher Co-Training Approach

Group members: Jingpei Wu, Ke Xin Chen, Kevin George

03.05.2021



Outline

- For the past two weeks
- Current problems and challenges
- Goals for the next two weeks

For the past two weeks

- Research and study
- Basic text preprocessing
  - Extracting relevant comments
  - Tokenizing, removing stop words (not sure), lemmatizing
- Started working on basic pre-trained processes
  - Using word2vec word embeddings in gensim
  - Using k-means for finding seed words (w stopwords and w/o stopwords)

Leveraging Just a Few Keywords for Fine-Grained Aspect Detection Through Weakly Supervised Co-Training, Karamanolakis et al.
Summarizing Opinions: Aspect Extraction Meets Sentiment Prediction and They Are Both Weakly Supervised, Angelidis et al.

Without stop words

1 [('fym', 21.496923), ('manuring', 23.954191), ('compact', 24.663473), ('hydraulic', 24.73682), ('insulation', 24.770617), ('sod', 25.120962), ('compaction', 25.623127), ('intercropping', 25.846485), ('savannah', 26.089266), ('portable', 26.210419), ('terrestrial', 26.555717), ('topography', 26.704634), ('damp', 26.711952), ('flushing', 27.018028), ('asphalt', 27.033289)]

2 [('pyridoxine', 10.493922), ('organophosphorus', 14.580263), ('toxicant', 16.234772), ('α', 17.260006), ('carbamate', 17.364964), ('phyto', 18.954506), ('halide', 19.249535), ('chelating', 19.351744), ('heptachlor', 19.947124), ('lyme', 20.41997), ('cobalamin', 20.54595), ('gliadin', 20.615366), ('malabsorption', 20.709042), ('enterotoxin', 21.60625), ('hardening', 21.629179)]

3 [('112', 23.960377), ('84', 25.9731), ('apr', 30.220932), ('280', 30.366135), ('73', 32.466072), ('51', 33.60796), ('92', 34.35273), ('m2', 34.681892), ('totaled', 36.3178), ('46', 38.937424), ('900', 39.046837), ('averaged', 39.43225), ('cagr', 39.498962), ('67', 40.069996), ('59', 40.823776)]

2 [('pyridoxine', 10.493922), ('organophosphorus', 14.580263), ('toxicant', 16.234772), ('α', 17.260006), ('carbamate', 17.364964), ('phyto', 18.954506), ('halide', 19.249535), ('chelating', 19.351744), ('heptachlor', 19.947124), ('lyme', 20.41997), ('cobalamin', 20.54595), ('gliadin', 20.615366), ('malabsorption', 20.709042), ('enterotoxin', 21.60625), ('hardening', 21.629179)]

chemical stuffs

6 [('ci_3299744', 0.82373774), ('emcphd', 2.1394513), ('2014g1062433', 2.316637), ('factorfizzle', 2.8161883), ('crediblehulk', 3.078110), ('noporganicstandards', 3.3338358), ('tfrec', 3.4121692), ('1003160', 3.6265867), ('0012346', 3.7127461), ('pmc3945755', 3.7866223), ('biolsci', 3.8057246), ('gl5gxarc134', 4.030347), ('gmfreecymru', 4.0466056), ('organic_certification', 4.2363653), ('pmc54831', 4.248015)]

7 [('department', 1.1368683e-13)]

8 [('galettes', 15.710376), ('creamy', 20.562258), ('crusty', 20.695404), ('stewed', 21.281387), ('toasted', 21.534874), ('beetroot', 21.556715), ('parsnip', 21.581305), ('unsalted', 22.0126), ('croissant', 23.00882), ('fennel', 23.311836), ('filet', 23.492228), ('tartare', 23.889305), ('fragrant', 23.943676), ('tomatoe', 24.341406), ('dumpling', 24.351759)]

9 [('1016', 6.086656), ('rajuktitus', 6.3434515), ('div6', 6.962177), ('gmoevidence', 7.0277205), ('as_vis', 7.0643163), ('entfacts', 7.1404347), ('1hom', 7.587775), ('linkclick', 7.7479186), ('panose', 8.043745), ('hess', 8.072983), ('ef130', 8.224717), ('3058', 8.314427), ('mohali', 8.844918), ('winn', 8.987744), ('prescott', 9.138864)]

With stop words

0 ['diversification', 'principally', 'restoration', 'intercropping', 'topography', 'transportability', 'agroforestry', 'fostering', 'estuary', 'offsetting', 'biodiverse', 'facilitating', 'monocrop', 'compaction', 'centralization']

1 ['countered', 'championed', 'infiltrated', 'applauded', 'graded', 'lectured', 'deployed', 'misinterpreted', 'alarmed', 'stimulated', 'vetted', 'alerted', 'hammered', 'tapped', 'ensured']

2 ['toasted', 'croissant', 'fluffy', 'parsnip', 'creamy', 'crusty', 'beetroot', 'aloe', 'unsalted', 'anchovy', 'fennel', 'dumpling', 'buttery', 'galettes', 'tomatillo']

3 ['foist', 'procure', 'broaden', 'discern', 'overwhelm', 'conceive', 'impress', 'fathom', 'align', 'empower', 'prioritize', 'decimate', 'accommodate', 'peddle', 'enact']

4 ['caricature', 'arrogantly', 'unhelpful', 'histrionics', 'zealous', 'naïve', 'nitpicking', 'conspiratorial', 'trite', 'intersting', 'belittling', 'uncritical', 'innuendo', 'nutcase', 'insinuation']

5 ['hess', 'enser', 'as_sdt', 'htdocs', 'arnold', 'linda', 'carter', 'ved', 'anderson', 'oakland', 'mj', 'wilson', 'oliver', 'leonard', 'psu']

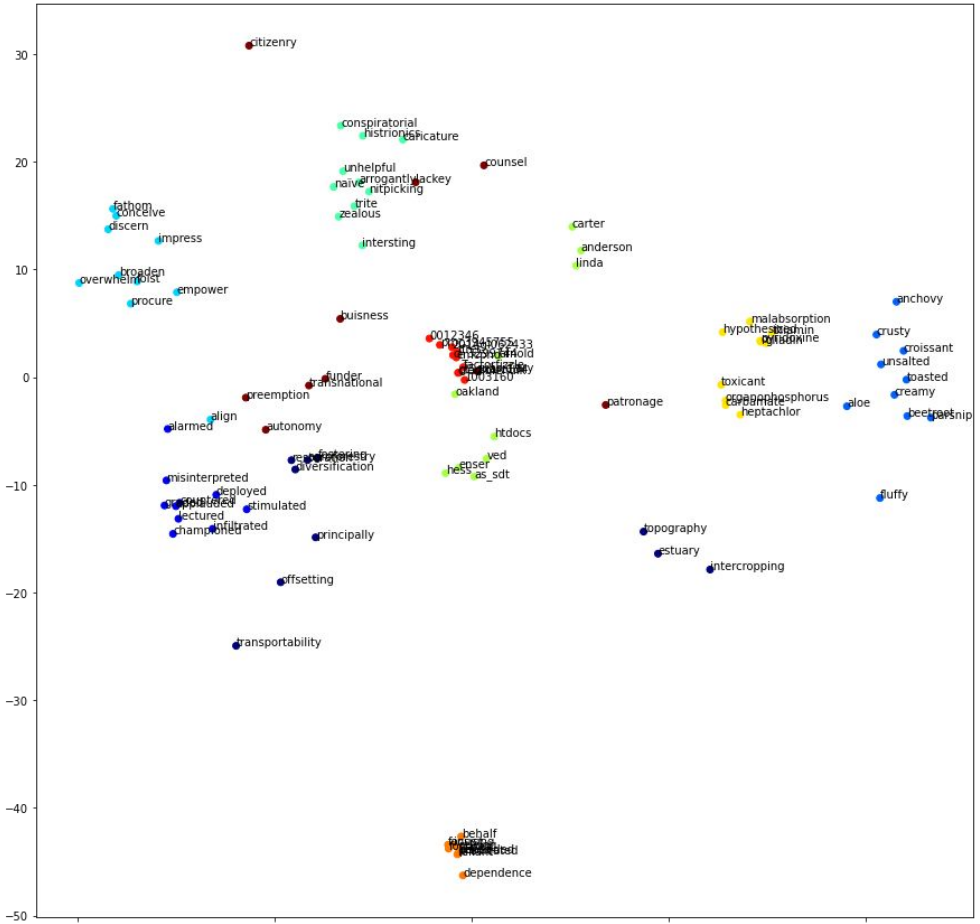
6 ['pyridoxine', 'toxicant', 'carbamate', 'α', 'gliadin', 'organophosphorus', 'thiamin', 'heptachlor', 'hypothesized', 'malabsorption', 'halide', 'benzoate', 'ester', 'sulfuric', 'solubility']

7 ['depended', 'predicated', 'reliant', 'relied', 'behalf', 'focusing', 'insist', 'focused', 'relying', 'dependence', 'reliance', 'concentrate', 'focus', 'relies', 'based']

8 ['ci_3299744', '2014g1062433', 'emcphd', 'tfrec', 'factorfizzle', 'g15gxarc134', '0012346', 'credible1hulk', '1003160', 'pmc3945755', 'p2566', '50331648', 'gmfreecymru', 'crucial24', 'charset']

9 ['lackey', 'counsel', 'funder', 'buisness', 'autonomy', 'preemption', 'transnational', 'patronage', 'subsidiary', 'citizenry', 'openness', 'dshea', 'profiteer', 'hierarchy', 'betterment']

t-SNE visualization



Current problem and challenges

- Collaboration on the code
- Reading official GitHub repo from [2] and unofficial code [3] for paper [1]
- Keep studying on certain methods and models (e.g Aspect Extraction method, NMF)

[1] Leveraging Just a Few Keywords for Fine-Grained Aspect Detection Through Weakly Supervised Co-Training, Karamanolakis et al.
[2] Summarizing Opinions: Aspect Extraction Meets Sentiment Prediction and They Are Both Weakly Supervised, Angelidis et al.
[3] https://github.com/aqweteddy/LeverageJustAFewKeywords

Goals for the next two weeks

- General schedule: setting up the whole pipeline with the simple word2vec and the important student-teacher co-training scheme, then add more settings, e.g. other word embedding models and comparisons
- Keep reading and understanding two GitHub repos for papers
- Get familiar with the annotated dataset
- Start applying NMF and clarity scoring function to get seed words
- Start building student-teacher modules

Q&A