Daniel Svoboda

NLP Research Scientist
at Johnson & Johnson

NLP SUMMIT

Online Event | Oct 4-6

Watch live:

Template Based Information Extraction with Dendograms to Classify News Articles

Problem – Classifying News Articles

– On a website, we want to automatically label a given news article when it arrives.

– It can become tedious to have to read through the article and figure out which category it belongs to.

– Both money resources (hiring employees) and time resources (browsing through the article) are wasted.

– Also, for medical news and research, what if they were collected into specific bins that would allow a researcher or doctor to easily find new insights?

Natural language processing can help

– NLP backed with machine learning has now made fabulous strides in classifying/clustering text.

– Classification is what is called supervised learning: Learning data with known labels.

– Clustering is what is called unsupervised learning. Inferring patterns from data where the labels are not known.

– Normal method is to use TF-IDF or word embeddings followed by classification.

Unsupervised clustering of articles

– Given that we have no idea of the classification of the articles, we will use unsupervised learning.

– Normal route is to get count vectorizer or tf-idf followed by dimension reduction and some form of clustering (usually k-means).

– Method only takes advantage of frequency of vocabulary and rarity of words.

– News articles of a given topic always have a given vocabulary and structure

– Linguistics of a technology article will be different from linguistics of a business article.

– What if there is a method that would help take advantage of that?

Implementation of Jurafsky's paper

– Matter of fact, there is a way to do so. An implementation can be found with the following paper: "Template-Based Information Extraction without the Templates" by Nathanel Chambers and Daniel Jurafsky.

– Paper discusses normal methods of extracting named entities and supervised learning as a way to extract entities that would classify document.

– It proposes an alternative approach that uses clustering of extracted "templates" and "semantic roles" that follow closely the linguistics of the article.

– Using "templates" and "semantic roles", we can infer the category of a document using clustering.

Templates from the paper

– Templates from the paper are defined as event patterns formed from a noun in Wordnet and a verb or a verb and head word of a syntactic object.

– Examples of this would be “explode”, “explosion”, etc. Linguistics seen here help to differentiate the topic relating to terrorism for example.

– Also to be extracted would be the named entity for a given person, thing, etc.

– Extracted entities for example in the term “killer”: person, bomb: thing, etc.

– Thus we would have ready several verbs, nouns and named entities if found for given verbs.

Slot filling

– With the extracted nouns, entities and verbs, we can perform slot filling in order to get the templates.

– Slot filling is the process of creating subcategories for each template.

– For example, for a bombing template, a bombing instrument is an object that explodes, is defused, etc.

– A bombing perpetrator is a person that detonates a bomb, sets off, is detained, etc.

– All these various slot filings are compatible with agglomerative clustering with the template being the overall cluster and the slots being the subclusters.

– Overall, extracted words plus distance measures are clustered with agglomerative clustering to find these templates in an unsupervised manner.

Example of templates.

Bombing Template (MUC-4)

Perpetrator *Person/Org* who detonates, blows up, plants, hurls, stages, is detained, is suspected, is blamed on, launches

Instrument A *physical object* that is exploded, explodes, is hurled, causes, goes off, is planted, damages, is set off, is defused

Target A *physical object* that is damaged, is destroyed, is exploded at, is damaged, is thrown at, is hit, is struck

Police *Person/Org* who raids, questions, discovers, investigates, defuses, arrests

N/A A *physical object* that is blown up, destroys

Attack/Shooting Template (MUC-4)

Perpetrator *Person/Org* who assassinates, patrols, ambushes, raids, shoots, is linked to

Victim *Person/Org* who is assassinated, is toppled, is gunned down, is executed, is evacuated

Target *Person/Org* who is hit, is struck, is downed, is set fire to, is blown up, surrounded

Instrument A *physical object* that is fired, injures, downs, is set off, is exploded

Kidnap Template (MUC-4)

Perpetrator *Person/Org* who releases, abducts, kidnaps, ambushes, holds, forces, captures, is imprisoned, frees

Target *Person/Org* who is kidnapped, is released, is freed, escapes, disappears, travels, is harmed, is threatened

Police *Person/Org* who rules out, negotiates, condemns, is pressured, finds, arrests, combs

Weapons Smuggling Template (NEW)

Perpetrator *Person/Org* who smuggles, is seized from, is captured, is detained

Police *Person/Org* who raids, seizes, captures, confiscates, detains, investigates

Instrument A *physical object* that is smuggled, is seized, is confiscated, is transported

Election Template (NEW)

Voter *Person/Org* who chooses, is intimidated, favors, is appealed to, turns out

Government *Person/Org* who authorizes, is chosen, blames, authorizes, denies

Candidate *Person/Org* who resigns, unites, advocates, manipulates, pledges, is blamed

Implementation of the program (Direct ideas)

– Based on the paper’s ideas, we can create a practical program for clustering.

– Get a count vectorizer of lemmatized verbs in the corpus.

– Get a count vectorizer of lemmatized verbs and dependencies (subject, object, etc.) from the corpus.

– Get count vectorizer of lemmatized verb and POS tag tuple pair for each document in the corpus.

– Get count vectorizer of two events (verb pairs) along with their dependency.

Implementation of the program (indirect ideas)

– Get a count vectorizer of POS tag count for each document in the corpus.

(Ideas is that different templates will have different distribution of POS tags.)

– Get a count vectorizer of entity count for each document in the corpus.

(Idea is that different templates will have different distributions of entities).

– Get ROBERTa sentence embedding of each document in a corpus.

(Idea is that important nouns and verbs will have higher attention score reflected in embeddings).

– Get topic modeling vector via Top2Vec.

(Idea is that topic modeling will create topics that will roughly sketch out the templates.)

Rationale for BERT and Top2Vec vectors

– Previous CountVectorizers collected statistics of words, entities, and dependencies in line with the paper.

– Additional ideas were to take advantages of deep learning and attention.

– BERT uses attention mechanisms to get embeddings for sentences.

– Attention likely will put more emphasis on words encompassing templates (nouns, verbs, etc.).

– Top2Vec creates a topic modeling vectors using deep learning to encode each topic as numerical vectors.

– Each document will have a vector reflecting mixture of topics. Aim is to find rough sketch of templates using topics as proxy.

Clustering of the data points

– Total number of variables usually are in the 10000's range. Curse of dimensionality problems.

– Experiments with various dimensionality reductions show UMAP is the best.

(It tends to preserve the topology of the data cloud into lower dimensions.)

– Perform dendogram clustering with a given number of clusters to reflect template numbers.

– By examining the cluster numbers, infer the template each document parses into, then the more abstract templates.

Example with CNN articles

– One example to show how well this system clusters with templates is by using CNN articles.

– 32 articles were selected at random from CNN’s website from 4 different categories: Style, Entertainment, Travel and Health.

– Program was run with UMAP set to 16 variables reducing from 13264.

– Aggolomorative Clustering was set to 4 clusters initially.

– Not all news articles were picked up correctly. Some were mistaken for other clusters.

– Aggolomorative Clustering was then increased to 8 clusters. Clusters were found to work perfectly with each new section clustering to 1 of 2 clusters.

Articles used from CNN

1. https://www.cnn.com/2022/09/09/health/abortion-restrictions-texas/index.html
2. https://www.cnn.com/2022/09/08/health/covid-kids-back-to-school/index.html
3. https://www.cnn.com/2022/09/08/health/omicron-booster-vaccine-health-wellness/index.html
4. https://www.cnn.com/2022/09/07/health/evusheld-antibodies-omicron-ba-4-6/index.html
5. https://www.cnn.com/2022/09/13/health/monkeypox-outbreak-improving-but-not-over/index.html
6. https://www.cnn.com/2022/09/14/health/daily-multivitamin-cognitive-function-study-wellness/index.html
7. https://www.cnn.com/2022/09/12/health/covid-booster-flu-shot-timing-explained-wellness/index.html
8. https://www.cnn.com/2022/09/08/health/fda-cancer-breast-implants/index.html
9. https://www.cnn.com/2022/09/11/entertainment/emmys-2022-preview/index.html
10. https://www.cnn.com/2022/09/10/entertainment/emmys-tv-filming-locations-cec/index.html
11. https://www.cnn.com/2022/09/10/entertainment/indiana-jones-d23-expo/index.html
12. https://www.cnn.com/2022/09/11/entertainment/monarch-review/index.html
13. https://www.cnn.com/2022/09/14/entertainment/the-handmaids-tale-season-5-review/index.html
14. https://www.cnn.com/2022/09/14/entertainment/pnb-rock-robbery-interview/index.html
15. https://www.cnn.com/2022/09/14/entertainment/tom-brady-gisele-bundchen/index.html
16. https://www.cnn.com/2022/09/14/entertainment/nielsen-report-latino-representation-tv-reaj/index.html
17. https://www.cnn.com/style/article/venice-film-festival-fashion-week-2-2022/index.html
18. https://www.cnn.com/style/article/how-to-shop-digital-fashion-september-issues/index.html
19. https://www.cnn.com/style/article/karlie-kloss-september-issues/index.html
20. https://www.cnn.com/style/article/jerry-seinfeld-models-kith-ltw/index.html
21. https://www.cnn.com/style/article/william-klein-fashion-photographer-death-tan/index.html
22. https://www.cnn.com/style/article/2022-emmys-red-carpet/index.html
23. https://www.cnn.com/style/article/fashion-metaverse-millions-september-issues/index.html
24. https://www.cnn.com/style/article/how-to-shop-digital-fashion-september-issues/index.html
25. https://www.cnn.com/travel/article/cdc-covid-travel-risk-destinations-august-29/index.html
26. https://www.cnn.com/travel/article/pumpkin-boat-record-attempt-nebraska/index.html
27. https://www.cnn.com/travel/article/air-travel-complaints-dot/index.html
28. https://www.cnn.com/travel/article/alaska-airlines-flight-diverted/index.html
29. https://www.cnn.com/travel/article/cdc-covid-travel-risk-destinations-september-12/index.html
30. https://www.cnn.com/travel/article/queen-elizabeth-raf-plane-most-tracked-flight-ever/index.html
31. https://www.cnn.com/travel/article/airline-passenger-prison-sentence-flight-crew-interference/index.html
32. https://www.cnn.com/travel/article/moenjodaro-pakistan-floods-intl-hnk/index.html

Results from Agglomerative Clustering for CNN articles

– Using 4 clusters, we got the following:

```
[0 0 0 0 0 0 0 0 1 0 1 1 1 1 1 1 1 2 2 1 1 1 2 2 3 3 3 3 3 3 3 3]
```

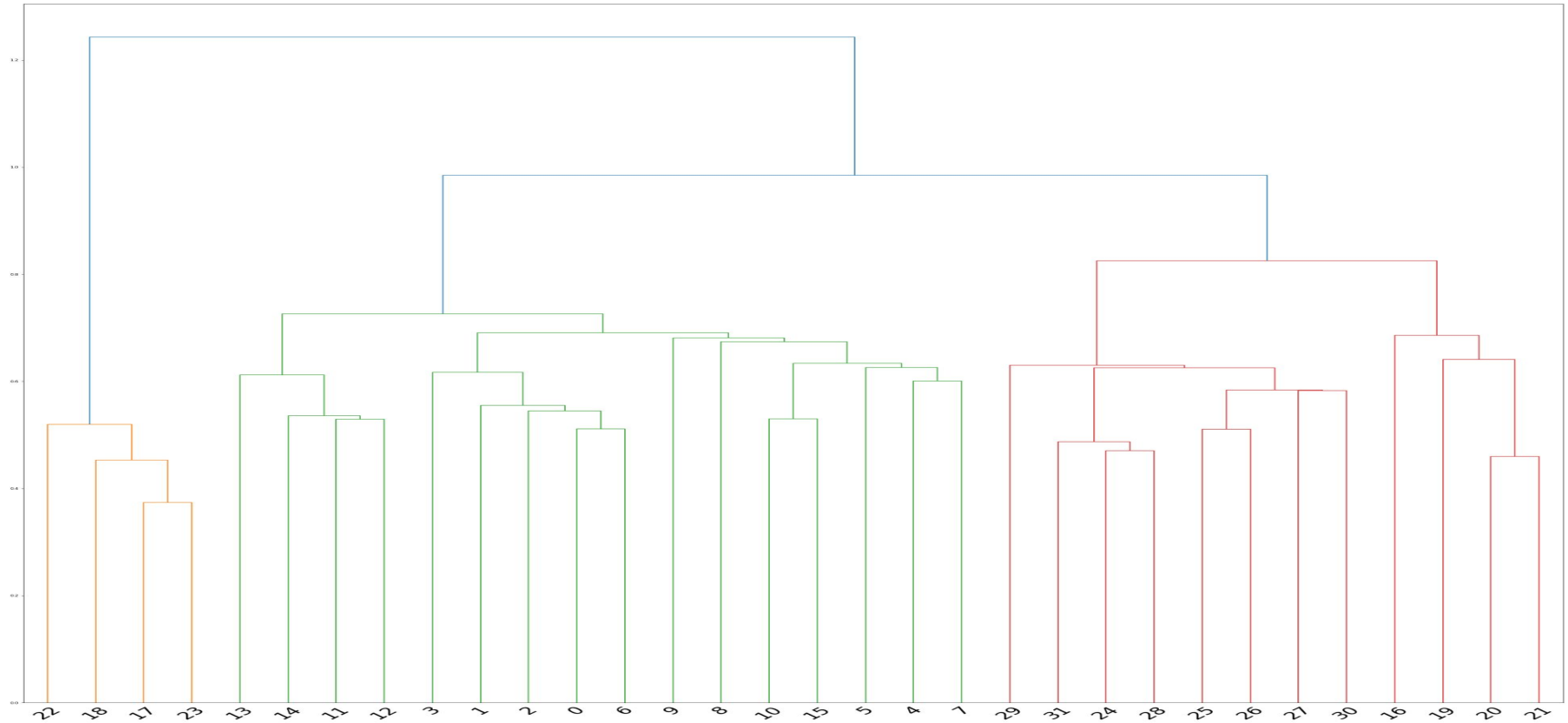
- All health are correct, entertainment has 1 error, style has 3 errors, travel are all correct.

– Using 8 clusters, we got the following:

```
[0 0 0 0 7 0 0 7 1 3 3 1 1 1 1 3 2 4 4 2 2 2 4 4 5 6 6 6 5 6 6 5]
```

– All are correct: Health are 0 and 7, entertainment are 3 and 1, style is 4 and 2, travel is 5 and 6.

Dendogram modeling of CNN articles



Interpretation of the Dendogram

– Style articles 22, 18, 17, 23 form their own unique template.

– Health and entertainment articles 0-15 have sub-templates (0,6 and 10,15) which all converge into a common template used by health and entertainment.

– Style and travel articles 16, 19-21, 24-31 all converge to a common template also.

– It would be interesting to see what templates health and entertainment converge to as a linguistic study.

– Also interesting for style and travel, although this is more intuitive.

– Using dendogram, one can cluster a future CNN article and get an idea of what template is being used for the clustering.

Example with cancer articles from cancer.gov

– The next example will demonstrate the template method for cancer news from cancer.gov.

– The website has a format of showing daily articles in an unstructured format.

– Articles deal with a mix of various topics: social issues, immunotherapies, gene editing, surgeries, recurrence of cancers, etc.

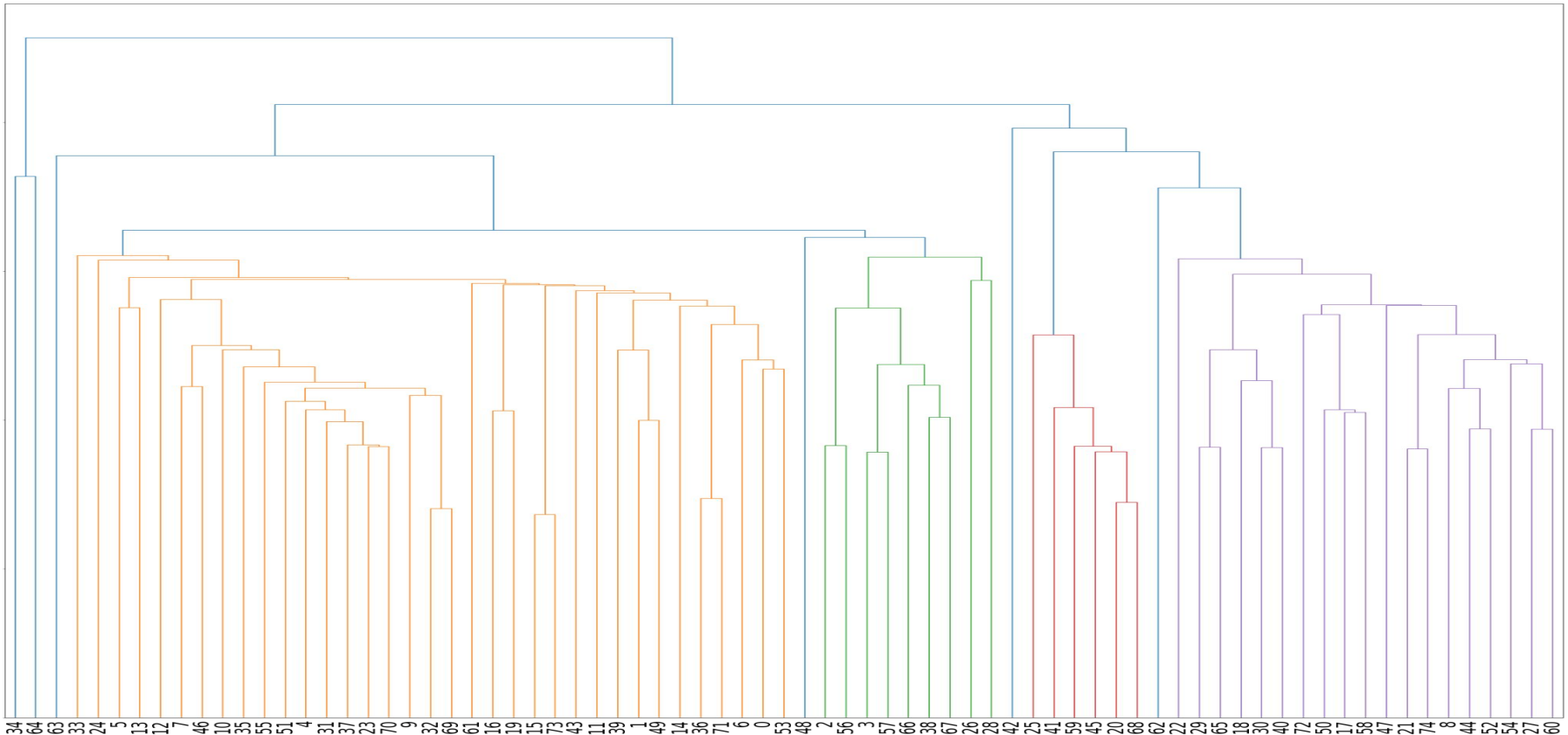
– Template method can automatically cluster articles into neater format.

– Even better, dendogram can yield insights as to sub-templates, thereby looking for connections in regards to research.

Articles used from cancer.gov

https://www.cancer.gov/news-events/cancer-currents-blog/2022/fda-lung-cancer-enhertu-her2
https://www.cancer.gov/news-events/cancer-currents-blog/2022/study-confirms-dinutiximab-high-risk-neuroblastoma
https://www.cancer.gov/news-events/cancer-currents-blog/2022/bladder-cancer-chemo-effective-alternative-bcg
https://www.cancer.gov/news-events/cancer-currents-blog/2022/liver-transplant-liver-cancer-downstaging
https://www.cancer.gov/news-events/cancer-currents-blog/2022/pancreatic-cancer-collagen-treatment-target
https://www.cancer.gov/news-events/cancer-currents-blog/2022/reducing-inflammation-to-treat-cancer
https://www.cancer.gov/news-events/cancer-currents-blog/2022/pembrolizumab-triple-negative-breast-cancer-improves-survival
https://www.cancer.gov/news-events/cancer-currents-blog/2022/rhabdomyosarcoma-targeting-fusion-protein-kdm4b
https://www.cancer.gov/news-events/cancer-currents-blog/2022/advanced-cancer-community-health-workers
https://www.cancer.gov/news-events/cancer-currents-blog/2022/micro-organospheres-cancer-model-treatment-response
https://www.cancer.gov/news-events/cancer-currents-blog/2022/cancer-immunotherapy-cmv-peptides
https://www.cancer.gov/news-events/cancer-currents-blog/2022/fda-dabrafenib-trametinib-braf-solid-tumors
https://www.cancer.gov/news-events/cancer-currents-blog/2022/melanoma-treatment-androgen-receptor
https://www.cancer.gov/news-events/cancer-currents-blog/2022/cancer-immunotherapy-skin-side-effects-microbes
https://www.cancer.gov/news-events/cancer-currents-blog/2022/ewing-sarcoma-chemotherapy-comparison
https://www.cancer.gov/news-events/cancer-currents-blog/2022/enhertu-her2-low-breast-cancer
https://www.cancer.gov/news-events/cancer-currents-blog/2022/glioma-brain-cancer-dabrafenib-trametinib
https://www.cancer.gov/news-events/cancer-currents-blog/2022/fda-proposes-rule-prohibiting-menthol-cigarettes
https://www.cancer.gov/news-events/cancer-currents-blog/2022/prostate-cancer-active-surveillance-increasing
https://www.cancer.gov/news-events/cancer-currents-blog/2022/meningioma-brain-tumor-new-classifications
https://www.cancer.gov/news-events/cancer-currents-blog/2022/new-online-june-2022
https://www.cancer.gov/news-events/cancer-currents-blog/2022/psychosocial-cancer-survivors-patricia-ganz
https://www.cancer.gov/news-events/cancer-currents-blog/2022/cervical-cancer-hpv-vaccine-one-dose-kenya
https://www.cancer.gov/news-events/cancer-currents-blog/2022/keto-bhb-prevent-colorectal-cancer
https://www.cancer.gov/news-events/cancer-currents-blog/2022/cancer-lymph-nodes-metastasis
https://www.cancer.gov/news-events/cancer-currents-blog/2022/treating-cancer-pregnancy-new-drugs
https://www.cancer.gov/news-events/cancer-currents-blog/2022/nivolumab-chemotherapy-neoadjuvant-lung-cancer
https://www.cancer.gov/news-events/cancer-currents-blog/2022/covid-increasing-cancer-screening
https://www.cancer.gov/news-events/cancer-currents-blog/2022/ivosidenib-chemotherapy-aml-idh1
https://www.cancer.gov/news-events/cancer-currents-blog/2022/positive-fit-stool-test-colonoscopy
https://www.cancer.gov/news-events/cancer-currents-blog/2022/skin-cancer-screening-melanoma-overdiagnosis
https://www.cancer.gov/news-events/cancer-currents-blog/2022/cancer-genetic-mutation-body-location
https://www.cancer.gov/news-events/cancer-currents-blog/2022/cancer-mosaic-mutations-embryo
https://www.cancer.gov/news-events/cancer-currents-blog/2022/finding-cancer-early-mced-tests
https://www.cancer.gov/news-events/cancer-currents-blog/2022/testicular-cancer-surgery-monitoring-mri-ct-scan
https://www.cancer.gov/news-events/cancer-currents-blog/2022/implanted-drug-factories-il2-ovarian-cancer
https://www.cancer.gov/news-events/cancer-currents-blog/2022/fda-opdualag-melanoma-lag-3
https://www.cancer.gov/news-events/cancer-currents-blog/2022/melanoma-brain-metastases-amyloid-beta
https://www.cancer.gov/news-events/cancer-currents-blog/2022/fda-carvykti-multiple-myeloma
https://www.cancer.gov/news-events/cancer-currents-blog/2022/darolutamide-survival-metastatic-prostate-cancer
https://www.cancer.gov/news-events/cancer-currents-blog/2022/artificial-intelligence-cancer-imaging
https://www.cancer.gov/news-events/cancer-currents-blog/2022/new-online-march-2022
https://www.cancer.gov/news-events/cancer-currents-blog/2022/cancer-treatment-women-severe-side-effects
https://www.cancer.gov/news-events/cancer-currents-blog/2022/trametinib-low-grade-serous-ovarian-cancer
https://www.cancer.gov/news-events/cancer-currents-blog/2022/pandemic-telehealth-surge-cancer-care
https://www.cancer.gov/news-events/cancer-currents-blog/2022/childhood-cancer-survivors-pregnancy-baby-health
https://www.cancer.gov/news-events/cancer-currents-blog/2022/medulloblastoma-nanoparticle-palbociclib-sapanisertib
https://www.cancer.gov/news-events/cancer-currents-blog/2022/overdue-cervical-cancer-screening-increasing
https://www.cancer.gov/news-events/cancer-currents-blog/2022/chronic-gvhd-naive-t-cell-depletion
https://www.cancer.gov/news-events/cancer-currents-blog/2022/biliary-tract-cancer-durvalumab-improves-survival
https://www.cancer.gov/news-events/cancer-currents-blog/2022/navajo-nation-commercial-tobacco-ban
https://www.cancer.gov/news-events/cancer-currents-blog/2022/spectrin-fatty-liver-disease-liver-cancer
https://www.cancer.gov/news-events/cancer-currents-blog/2022/cancer-screening-presidents-cancer-panel-report
https://www.cancer.gov/news-events/cancer-currents-blog/2022/immunotherapy-cancer-biomarker-hla-gene
https://www.cancer.gov/news-events/cancer-currents-blog/2022/financial-problems-advanced-cancer
https://www.cancer.gov/news-events/cancer-currents-blog/2022/mrna-vaccines-to-treat-cancer
https://www.cancer.gov/news-events/cancer-currents-blog/2022/nhl-car-t-cells-belinda-transform-zuma7
https://www.cancer.gov/news-events/cancer-currents-blog/2022/ovarian-cancer-return-surgery-desktop-iii
https://www.cancer.gov/news-events/cancer-currents-blog/2022/ecigarettes-zeller-fda-regulation
https://www.cancer.gov/news-events/cancer-currents-blog/2021/new-online-december-2021
https://www.cancer.gov/news-events/cancer-currents-blog/2021/black-white-cancer-disparities-survival-accure
https://www.cancer.gov/news-events/cancer-currents-blog/2021/advanced-melanoma-braf-immunotherapy-first
https://www.cancer.gov/news-events/cancer-currents-blog/2021/cancer-geriatric-assessment-fewer-side-effects
https://www.cancer.gov/news-events/cancer-currents-blog/2021/breast-cancer-ddr1-collagen-barrier
https://www.cancer.gov/news-events/cancer-currents-blog/2021/medulloblastoma-children-test-residual-disease
https://www.cancer.gov/news-events/cancer-currents-blog/2021/breast-cancer-risk-calculator-us-black-women
https://www.cancer.gov/news-events/cancer-currents-blog/2021/prostate-cancer-hypoprt-radiation-therapy-safe
https://www.cancer.gov/news-events/cancer-currents-blog/2021/fda-tecartus-adults-b-cell-all
https://www.cancer.gov/news-events/cancer-currents-blog/2021/new-online-november-2021
https://www.cancer.gov/news-events/cancer-currents-blog/2021/protein-interactions-mapping-cancer-pathways
https://www.cancer.gov/news-events/cancer-currents-blog/2021/fructose-promotes-obesity-colorectal-cancer
https://www.cancer.gov/news-events/cancer-currents-blog/2021/fda-adjuvant-atezolizumab-lung-cancer
https://www.cancer.gov/news-events/cancer-currents-blog/2021/hpv-vaccine-parents-safety-concerns
https://www.cancer.gov/news-events/cancer-currents-blog/2021/enhertu-her2-metastatic-breast-cancer
https://www.cancer.gov/news-events/cancer-currents-blog/2021/living-with-metastatic-cancer

Dendogram for articles from cancer.gov



Analysis of the dendogram

– One bunch is 34, 64, and 63. (Group 1)

– Another bunch is 2, 56, 3, 57, 66, 38, 67, 26 and 28. (Group 2)

– Another bunch is 33, 24, 5, 13, 12, 7, 46, 10, 35, 55, 51, 4, 31, 37, 23, 70, 9, 32, 69, 61, 16, 19, 15, 73, 43, 11, 39, 1, 49, 14, 36, 71, 6, 0, 53 (Group 3)

– Another bunch is 25, 41, 59, 45, 20, and 68 (Group 4)

– Last bunch is 22, 29, 65, 18, 30, 40, 72, 50, 17, 58, 47, 21, 74, 8, 44, 52, 54, 27 and 60 (Group 5)

Results of the analysis of the dendogram

– Group 1 deals with causes for resistance to cancer treatments along with causes for recurrence after operations and/or therapy.

– Group 2 deals with initial or secondary methods that successfully inhibit or prevent recurrence of cancers. Each dendogram cluster focuses on a finer combination of one or two methods.

– Group 3 deals with solely immunotherapy treatments of various cancers.

– Group 4 deals with monthly roundup of articles.

– Group 5 deals with social issues involved with cancer (discrimination in testing, financial problems, telehealth, etc.)

Examples of finer granularity (53, 0, 6, 71 and 36)

– Document 53 and 0 deal with a very fine granularity of immunotherapy solutions based on specific DNA proteins in cancer cells.

  – https://www.cancer.gov/news-events/cancer-currents-blog/2022/immunotherapy-cancer-biomarker-hla-gene

  – https://www.cancer.gov/news-events/cancer-currents-blog/2022/fda-lung-cancer-enhertu-her2

– Document 6 deals with using immunotherapy solutions along with chemotherapy.

  – https://www.cancer.gov/news-events/cancer-currents-blog/2022/pembrolizumab-triple-negative-breast-cancer-improves-survival

– Documents 53, 0, and 6 all go back to a higher cluster of the dendogram. Here, they form a sub-template dealing with immunotherapy solutions to various forms of cancer.

– Documents 71 and 36 deal with use of immunotherapy drugs for specific cases of lung cancer.

  – https://www.cancer.gov/news-events/cancer-currents-blog/2022/pembrolizumab-triple-negative-breast-cancer-improves-survival

  – https://www.cancer.gov/news-events/cancer-currents-blog/2022/fda-opdualag-melanoma-lag-3

– These documents cluster with the subcluster formed from 53, 0 and 6. Together, they make a cluster related to more broader immunotherapy solutions.

– Documents 53 and 0 deal with immunotherapy solutions based on specific DNA proteins; could researcher yield insights into newer forms of cancer therapy?

Examples of finer granularity (69, 32, and 9)

– Documents 69 and 32 deal with effects of protein molecules related to ketosis (ketones and collagen) in inhibiting various forms of cancer.

  – https://www.cancer.gov/news-events/cancer-currents-blog/2021/protein-interactions-mapping-cancer-pathways

  – https://www.cancer.gov/news-events/cancer-currents-blog/2022/keto-bhb-prevent-colorectal-cancer

– Document 9 relates to protein molecules found in Alzheimer's that could potentially spread melanoma to the brain.

  – https://www.cancer.gov/news-events/cancer-currents-blog/2022/melanoma-brain-metastases-amyloid-beta

– Overall the subcluster relates to how protein molecules can either spread or inhibit various forms of cancer.

– For a small sample (75 documents), algorithm has clustered 69 and 32 into finer granularity of ketosis.

– IT Developer can create specific category for researchers. More importantly, researchers can analyze specific category to gather new insights in their research field, especially by analyzing the linguistics/templates.

Examples of finer granularity (68, 39, 67)

– Documents 68 and 39 deal with using CAR T-cell therapy as secondary treatments following initial therapy to inhibit spread of cancer.

  – https://www.cancer.gov/news-events/cancer-currents-blog/2021/fda-tecartus-adults-b-cell-all

  – https://www.cancer.gov/news-events/cancer-currents-blog/2022/fda-carvykti-multiple-myeloma

– Document 67 deals with using more intense radiation therapy after surgery for prostate cancer as a successful secondary treatment.

  – https://www.cancer.gov/news-events/cancer-currents-blog/2022/fda-carvykti-multiple-myeloma

– Both items cluster into a sub-cluster that emphasizes successful secondary treatments after initial ones.

– Documents 57 and 3 deal with secondary surgeries following initial ones to successfully prevent recurrences.

  – https://www.cancer.gov/news-events/cancer-currents-blog/2022/ovarian-cancer-return-surgery-desktop-iii

  – https://www.cancer.gov/news-events/cancer-currents-blog/2022/liver-transplant-liver-cancer-downstaging

– Both documents cluster into a higher cluster with 68, 39, and 67 that emphasize secondary treatments that are successful.

Conclusions

– Template method with dendograms helps to cluster news and potentially research articles into very fine subdomains.

– Subdomains can be discovered by examining articles clustered into dendogram or examining the entities, POS tags, verb pairs, etc.

– News or research sites can simply get classification through method instead of visually examining article.

– Researchers can visually examine templates and slot domains to discover interesting insights for news/research articles.

– Researchers could potentially take health articles/news and find within a fine cluster domain. Shuffling papers may potentially allow researchers to come up with novel discoveries and insights.

Questions?