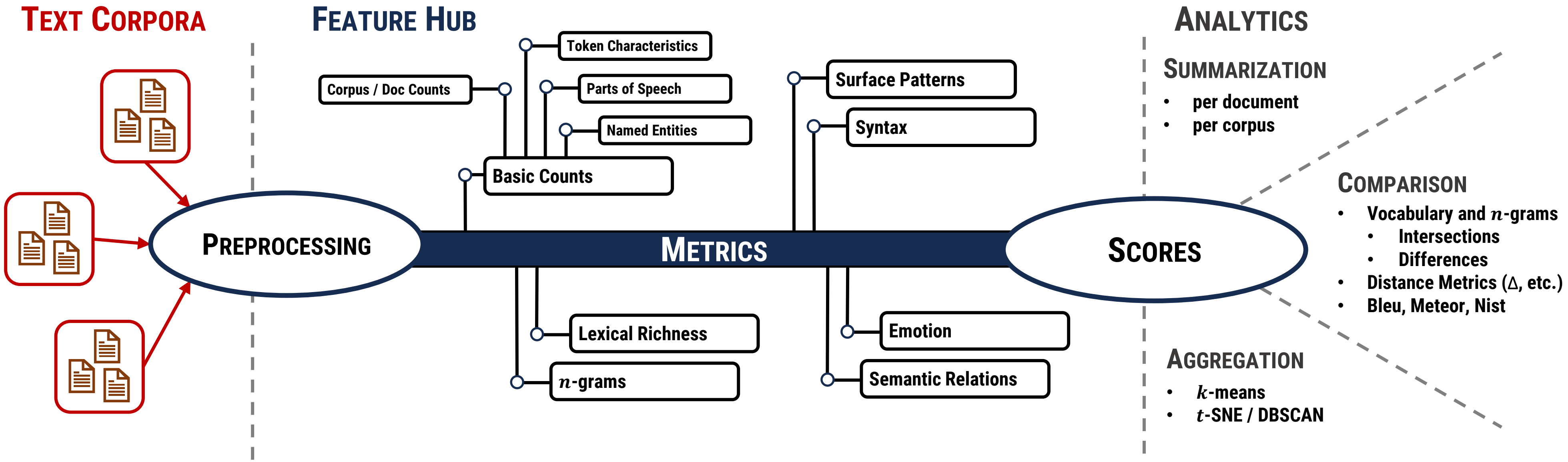
DoPa Meter -
A Tool Suite for Metrical Document Profiling and Aggregation

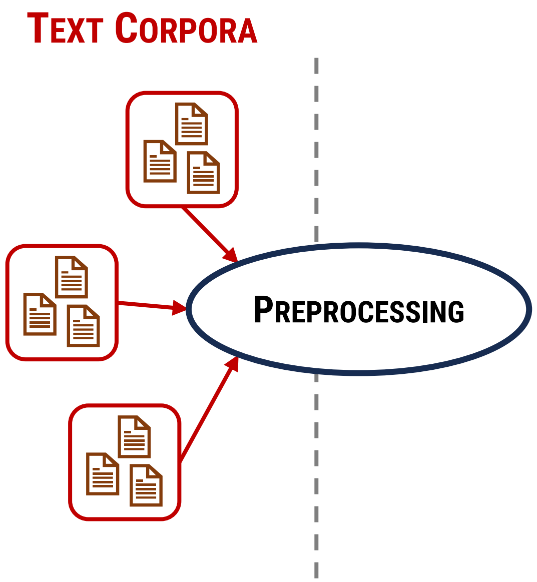


Summarization
- per document
- per corpus

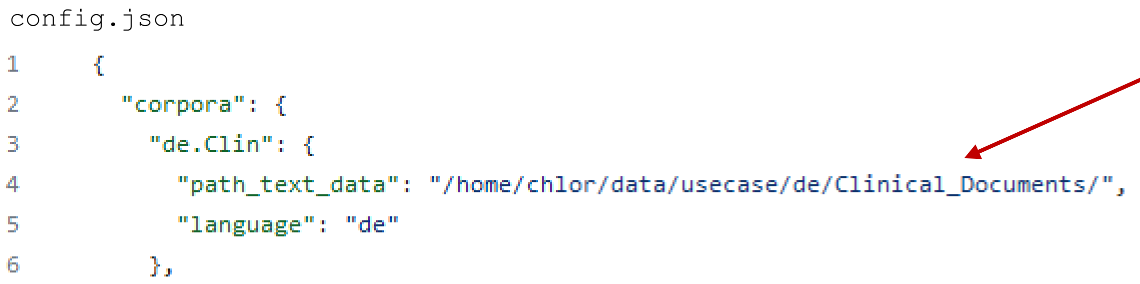
Comparison
- Vocabulary and n-grams
  - Intersections
  - Differences
- Distance Metrics (Δ, etc.)
- Bleu, Meteor, Nist

Aggregation
- k-means
- t-SNE / DBSCAN

Input: Corpora

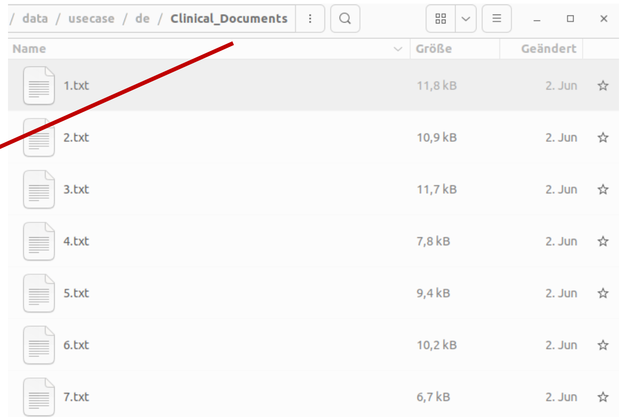


- Input: set of corpora
- 1 corpus: 1 directory of single files of plain text
- Preprocessing by spaCy



config.json

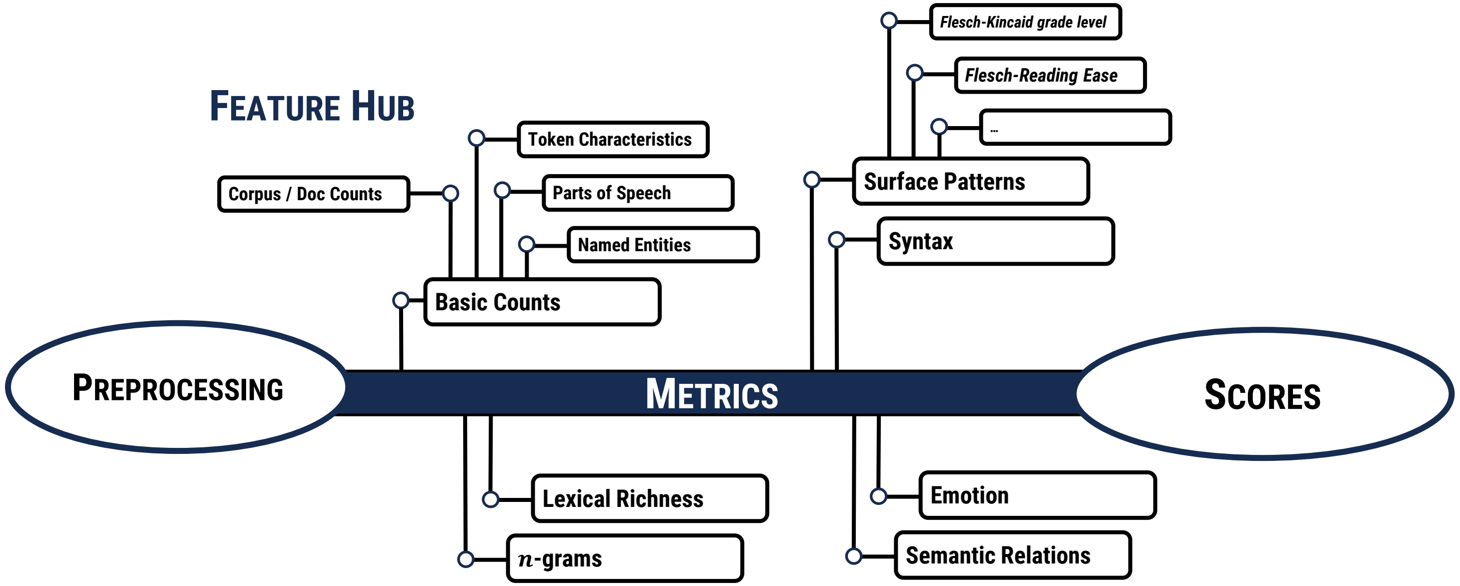
```
{
  "corpora": {
    "de.Clin": {
      "path_text_data": "/home/chlor/data/usecase/de/Clinical_Documents/",
      "language": "de"
    },
```



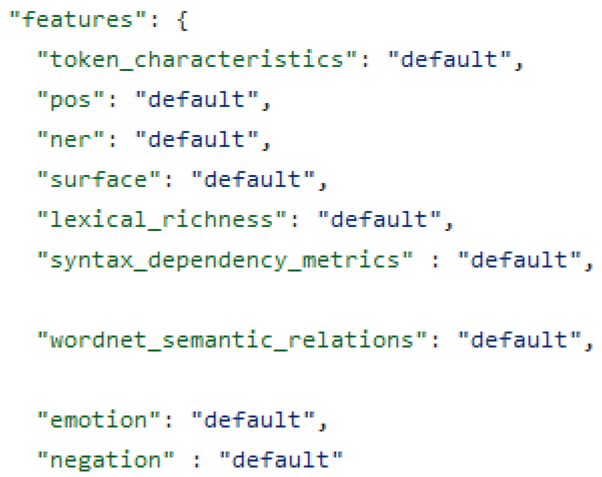
/ data / usecase / de / Clinical_Documents

Name	Größe	Geändert	
1.txt	11,8 kB	2. Jun	☆
2.txt	10,9 kB	2. Jun	☆
3.txt	11,7 kB	2. Jun	☆
4.txt	7,8 kB	2. Jun	☆
5.txt	9,4 kB	2. Jun	☆
6.txt	10,2 kB	2. Jun	☆
7.txt	6,7 kB	2. Jun	☆

Feature Hub



- Sets of single features
- Computation of features allows for an individual mode or a default mode, that computes all features.

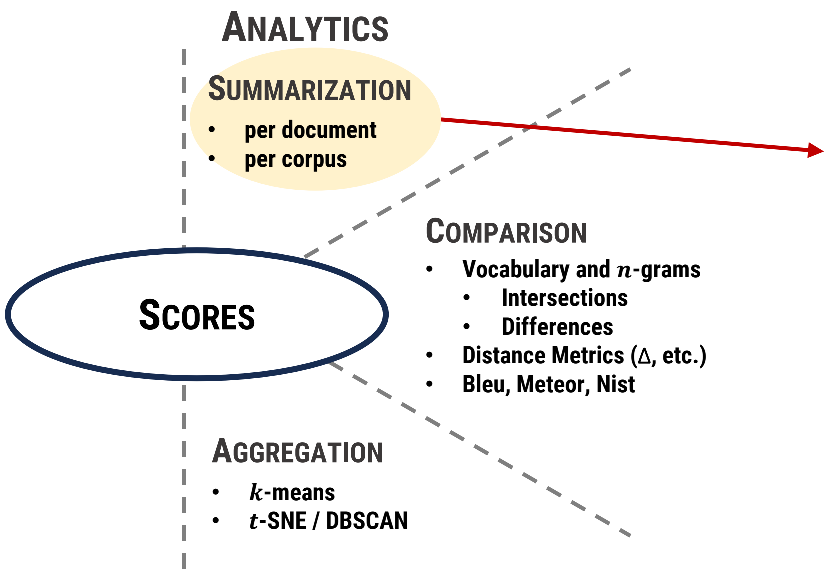


```
"features": {
  "token_characteristics": "default",
  "pos": "default",
  "ner": "default",
  "surface": "default",
  "lexical_richness": "default",
  "syntax_dependency_metrics" : "default",

  "wordnet_semantic_relations": "default",

  "emotion": "default",
  "negation" : "default"
```

More details under

- https://github.com/JULIELab/dopameter/tree/main/doc/features
- https://github.com/JULIELab/dopameter/tree/main/doc/res/example_configurations

Analytics - Summarization



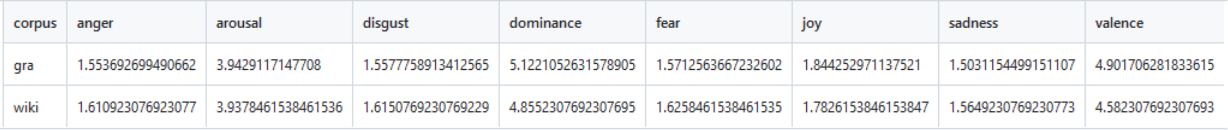
ANALYTICS

SUMMARIZATION
- per document
- per corpus

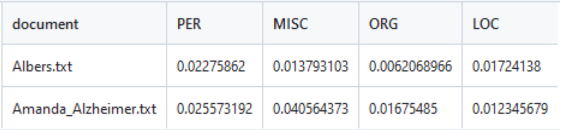
COMPARISON
- Vocabulary and n-grams
  - Intersections
  - Differences
- Distance Metrics (Δ, etc.)
- Bleu, Meteor, Nist

SCORES

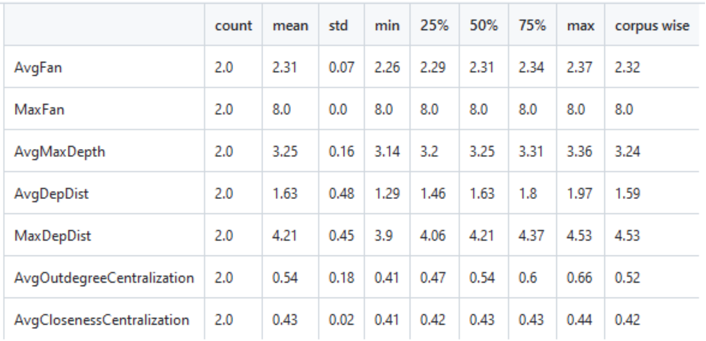
AGGREGATION
- k-means
- t-SNE / DBSCAN



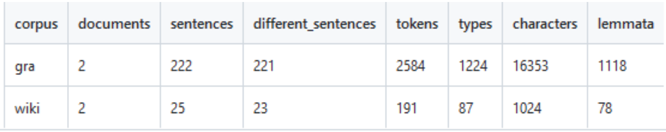
corpus	anger	arousal	disgust	dominance	fear	joy	sadness	valence
gra	1.553692699490662	3.9429117147708	1.5577758913412565	5.1221052631578905	1.5712563667232602	1.844252971137521	1.5031154499151107	4.901706281833615
wiki	1.610923076923077	3.9378461538461536	1.6150769230769229	4.8552307692307695	1.6258461538461535	1.7826153846153847	1.5649230769230773	4.582307692307693



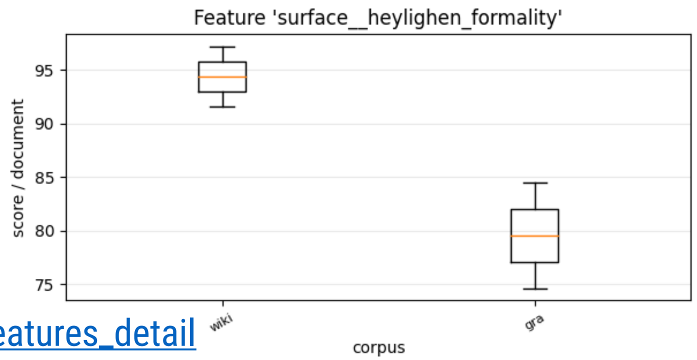
document	PER	MISC	ORG	LOC
Albers.txt	0.02275862	0.013793103	0.0062068966	0.01724138
Amanda_Alzheimer.txt	0.025573192	0.040564373	0.01675485	0.012345679



	count	mean	std	min	25%	50%	75%	max	corpus wise
AvgFan	2.0	2.31	0.07	2.26	2.29	2.31	2.34	2.37	2.32
MaxFan	2.0	8.0	0.0	8.0	8.0	8.0	8.0	8.0	8.0
AvgMaxDepth	2.0	3.25	0.16	3.14	3.2	3.25	3.31	3.36	3.24
AvgDepDist	2.0	1.63	0.48	1.29	1.46	1.63	1.8	1.97	1.59
MaxDepDist	2.0	4.21	0.45	3.9	4.06	4.21	4.37	4.53	4.53
AvgOutdegreeCentralization	2.0	0.54	0.18	0.41	0.47	0.54	0.6	0.66	0.52
AvgClosenessCentralization	2.0	0.43	0.02	0.41	0.42	0.43	0.43	0.44	0.42



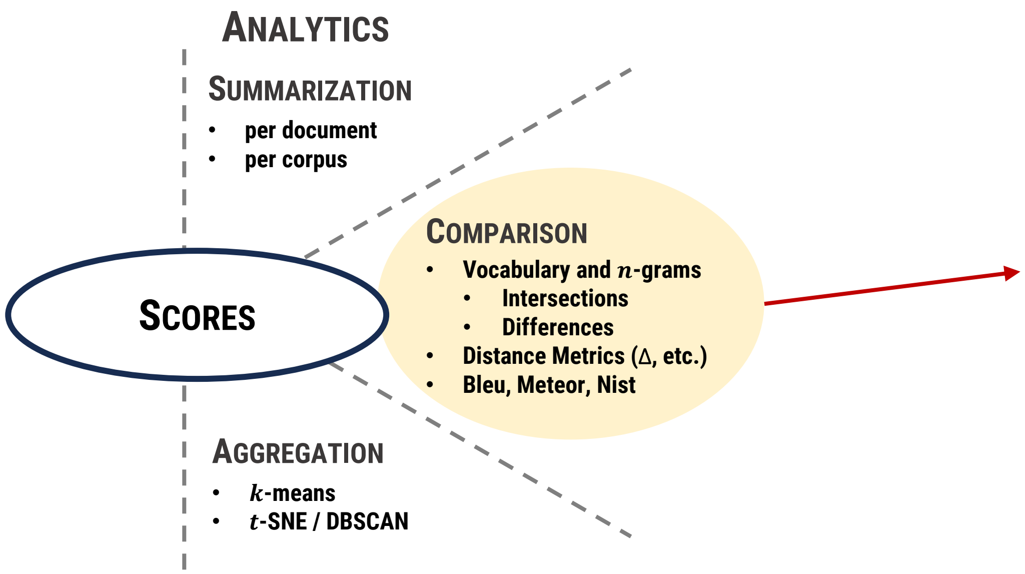
corpus	documents	sentences	different_sentences	tokens	types	characters	lemmata
gra	2	222	221	2584	1224	16353	1118
wiki	2	25	23	191	87	1024	78



Feature 'surface__heylighen_formality'

Examples under
- https://github.com/JULIELab/dopameter/tree/main/doc/res/results/features_detail
- https://github.com/JULIELab/dopameter/tree/main/doc/res/results/summary

Analytics - Comparison



Analytics

Summarization

- per document
- per corpus

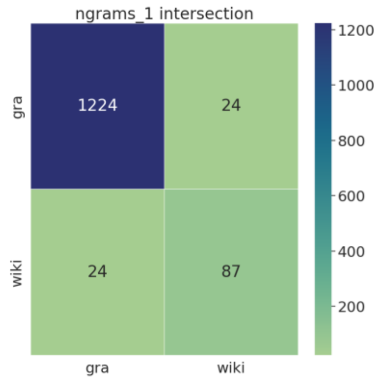
Scores

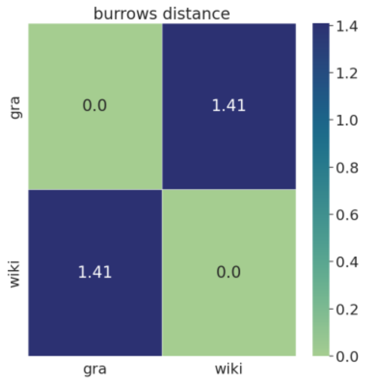
Comparison

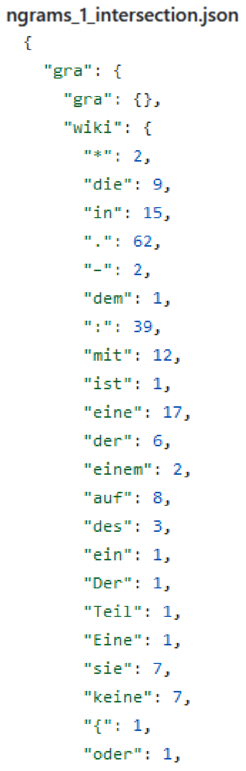
- Vocabulary and n-grams
  - Intersections
  - Differences
- Distance Metrics (Δ, etc.)
- Bleu, Meteor, Nist

Aggregation

- k-means
- t-SNE / DBSCAN







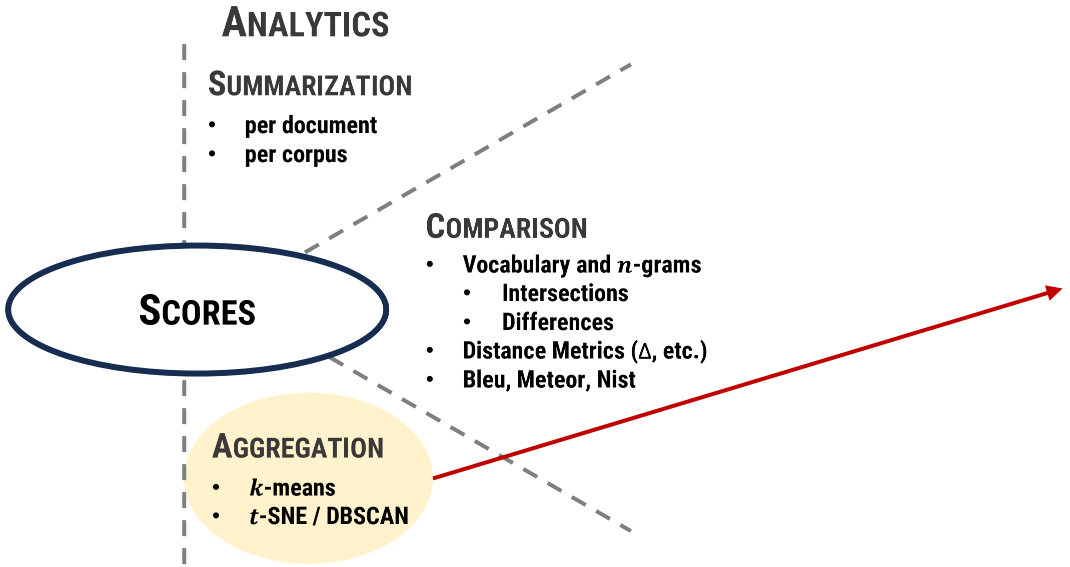
ngrams_1_intersection.json

```
{
  "gra": {
    "gra": {},
    "wiki": {
      "*": 2,
      "die": 9,
      "in": 15,
      ".": 62,
      "-": 2,
      "dem": 1,
      ":": 39,
      "mit": 12,
      "ist": 1,
      "eine": 17,
      "der": 6,
      "einem": 2,
      "auf": 8,
      "des": 3,
      "ein": 1,
      "Der": 1,
      "Teil": 1,
      "Eine": 1,
      "sie": 7,
      "keine": 7,
      "{": 1,
      "oder": 1,
```

Examples under

- https://github.com/JULIELab/dopameter/tree/main/doc/res/results/compare

Analytics - Aggregation



Analytics

Summarization

- per document
- per corpus

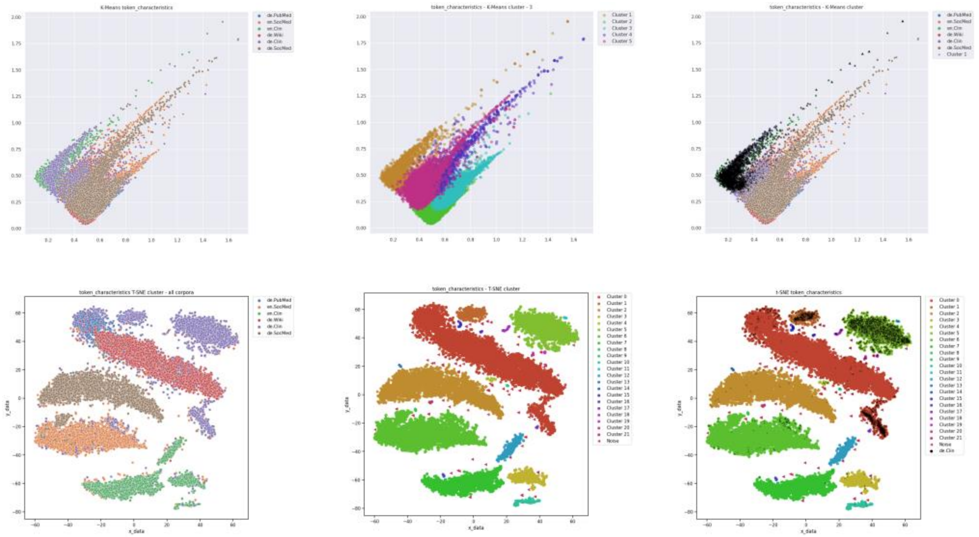
Comparison

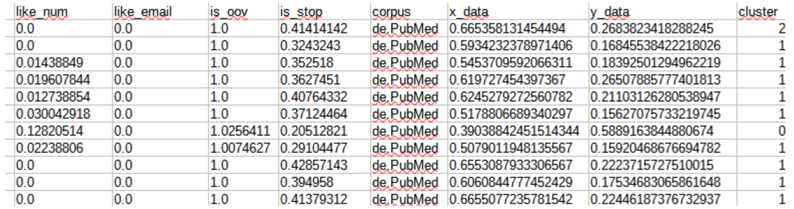
- Vocabulary and n-grams
  - Intersections
  - Differences
- Distance Metrics (Δ, etc.)
- Bleu, Meteor, Nist

Scores

Aggregation

- k-means
- t-SNE / DBSCAN



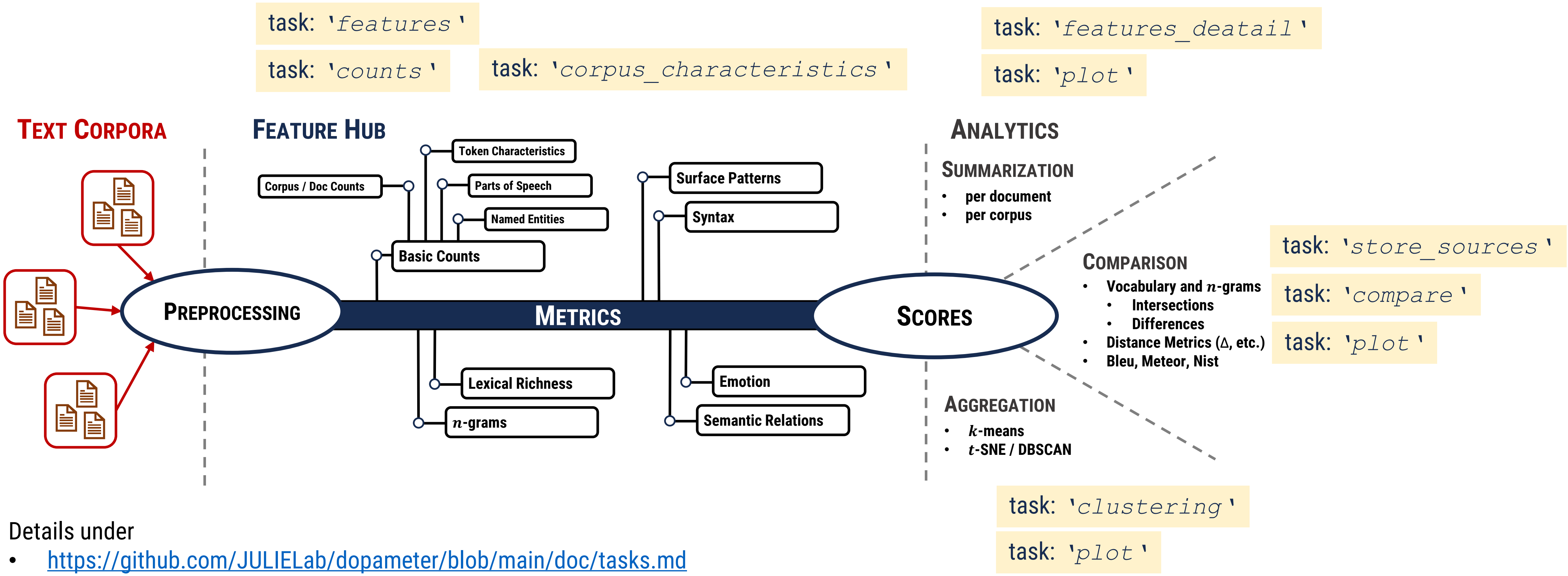


like_num	like_email	is_oov	is_stop	corpus	x_data	y_data	cluster
0.0	0.0	1.0	0.41414142	de.PubMed	0.665358131454494	0.2683823418288245	2
0.0	0.0	1.0	0.3243243	de.PubMed	0.5934232378971406	0.16845538422218026	1
0.01438849	0.0	1.0	0.352518	de.PubMed	0.5453709592066311	0.18392501294962219	1
0.019607844	0.0	1.0	0.3627451	de.PubMed	0.619727454397367	0.26507885777401813	1
0.012738854	0.0	1.0	0.40764332	de.PubMed	0.6245279272560782	0.21103126280538947	1
0.030042918	0.0	1.0	0.37124464	de.PubMed	0.5178806689340297	0.15627075733219745	1
0.12820514	0.0	1.0256411	0.20512821	de.PubMed	0.39038842451514344	0.5889163844880674	0
0.02238806	0.0	1.0074627	0.29104477	de.PubMed	0.5079011948135567	0.15920468676694782	1
0.0	0.0	1.0	0.42857143	de.PubMed	0.6553087933306567	0.2223715727510015	1
0.0	0.0	1.0	0.394958	de.PubMed	0.6060844777452429	0.17534683065861648	1
0.0	0.0	1.0	0.41379312	de.PubMed	0.6655077235781542	0.22446187376732937	1

Examples under

- https://github.com/JULIELab/dopameter/tree/main/doc/res/example_aggregation

Technical Notes for configuration and running



task: 'features'

task: 'counts'

task: 'corpus_characteristics'

task: 'features_deatail'

task: 'plot'

task: 'store_sources'

task: 'compare'

task: 'plot'

task: 'clustering'

task: 'plot'

Details under

- https://github.com/JULIELab/dopameter/blob/main/doc/tasks.md

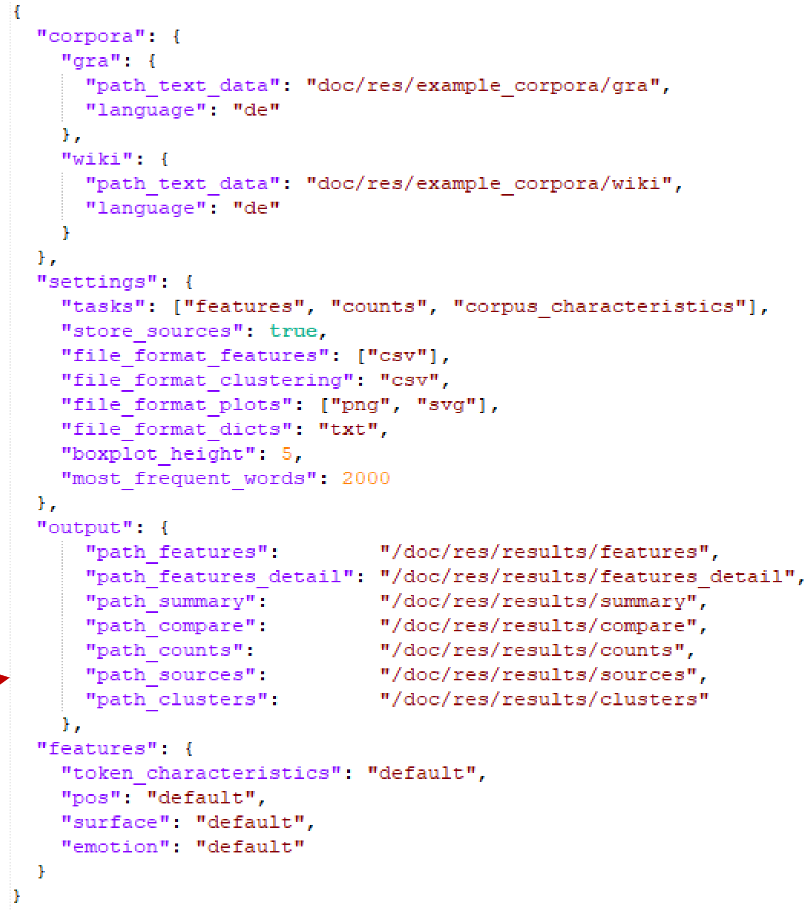
Try DoPa Meter

- Install Python 3

- Installation external sources:
  - `python install_languages.py lang_install.json`

- Define `config.json`

- Open a terminal or PowerShell under Windows and Run
  - `python main.py config.json`



```
{
  "corpora": {
    "gra": {
      "path_text_data": "doc/res/example_corpora/gra",
      "language": "de"
    },
    "wiki": {
      "path_text_data": "doc/res/example_corpora/wiki",
      "language": "de"
    }
  },
  "settings": {
    "tasks": ["features", "counts", "corpus_characteristics"],
    "store_sources": true,
    "file_format_features": ["csv"],
    "file_format_clustering": "csv",
    "file_format_plots": ["png", "svg"],
    "file_format_dicts": "txt",
    "boxplot_height": 5,
    "most_frequent_words": 2000
  },
  "output": {
      "path_features":        "/doc/res/results/features",
      "path_features_detail": "/doc/res/results/features_detail",
      "path_summary":         "/doc/res/results/summary",
      "path_compare":         "/doc/res/results/compare",
      "path_counts":          "/doc/res/results/counts",
      "path_sources":         "/doc/res/results/sources",
      "path_clusters":        "/doc/res/results/clusters"
    },
  "features": {
    "token_characteristics": "default",
    "pos": "default",
    "surface": "default",
    "emotion": "default"
  }
}
```