Topic Modeling Open NASA Data

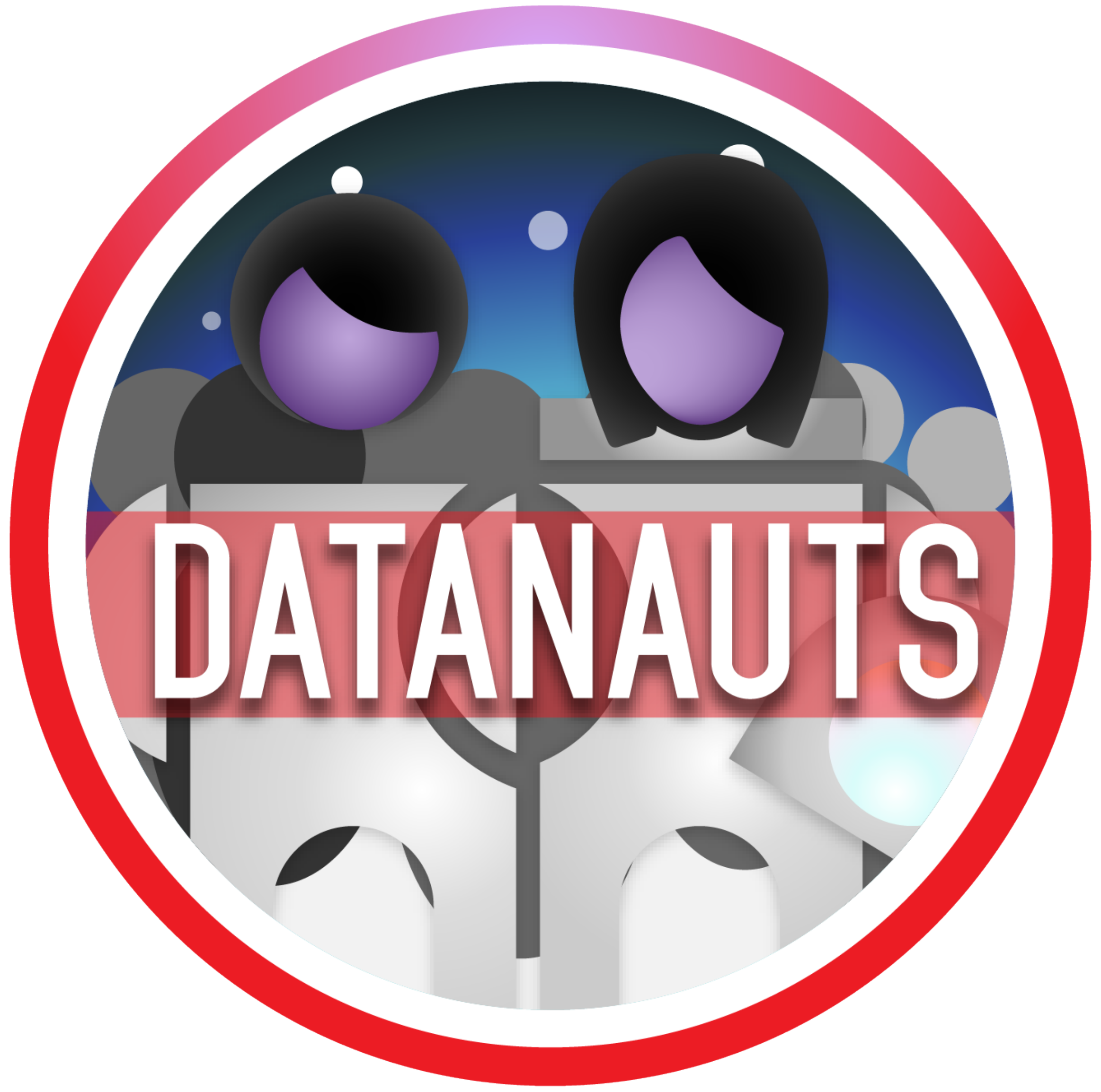
Noemi Derzsy

Strata
DATA CONFERENCE

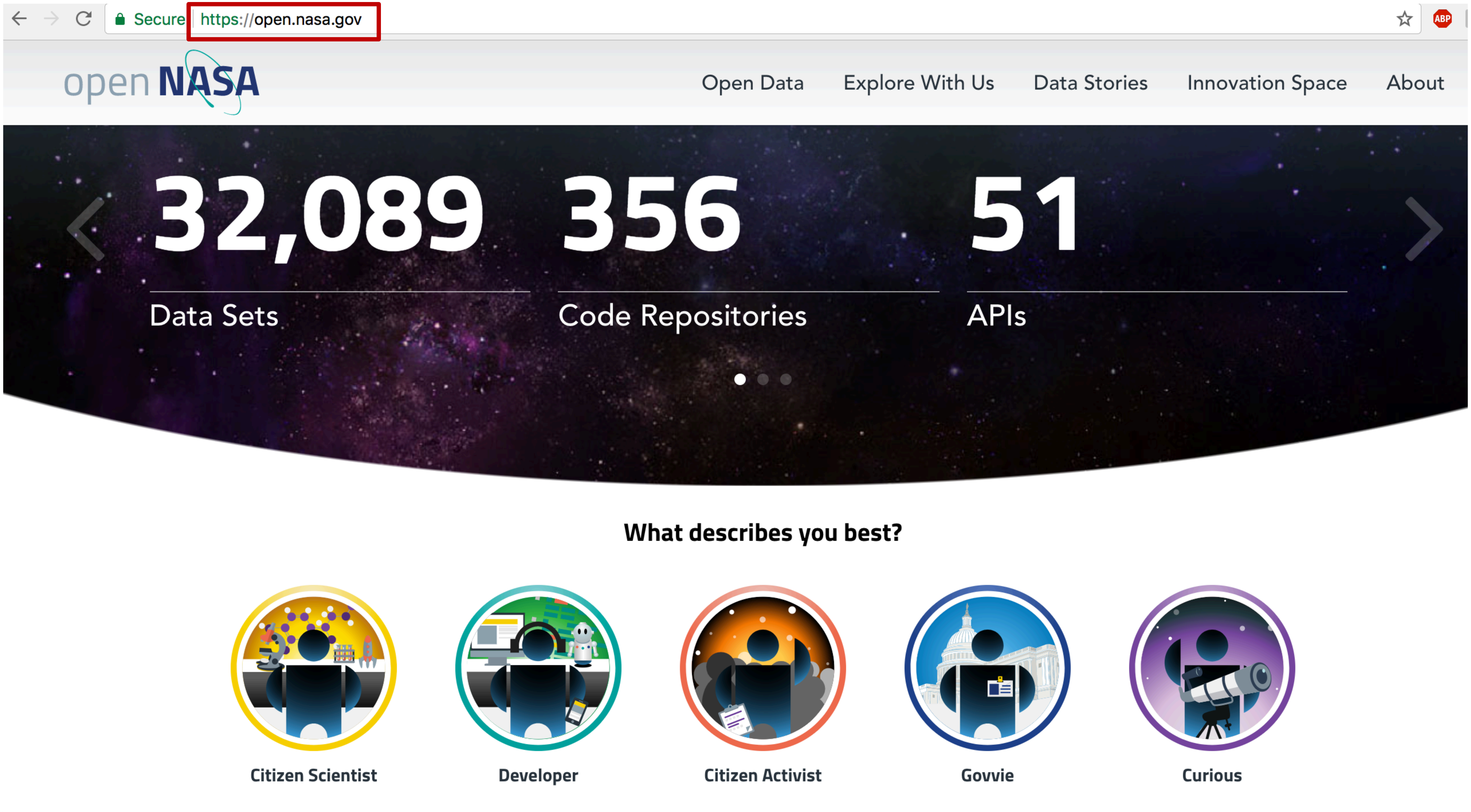
Open NASA Platform

- Data at NASA: 32 089
- Data at other government agencies: ~185 000
- NASA code repositories: 356
- NASA APIs: 51

Continuously growing…



https://open.nasa.gov/explore/datanauts/


Secure
https://open.nasa.gov
open NASA
Open Data
Explore With Us
Data Stories
Innovation Space
About
32,089
Data Sets
356
Code Repositories
51
APIs
What describes you best?
Citizen Scientist
Developer
Citizen Activist
Govvie
Curious

Open NASA Metadata

- Datasets: 32089
- Some of the data sets:
  - Mars Rover sound data
  - Hubble Telescope image collection
  - NASA patents
  - Picture of the Day of Earth
  - etc.

http://data.nasa.gov/data.json

Metadata information:

- id
- type
- accessLevel
- accrualPeriodicity
- bureauCode
- contactPoint
- title
- description
- distribution
- identifier
- Issued
- keyword
- language
- modified
- programCode
- theme
- license
- location (HTML link)
- Etc.

Format: JSON

Which of these features is "best" to tie together the data?

How do we label groupings in a meaningful manner?

How many groups/how to arrange/visualize them?

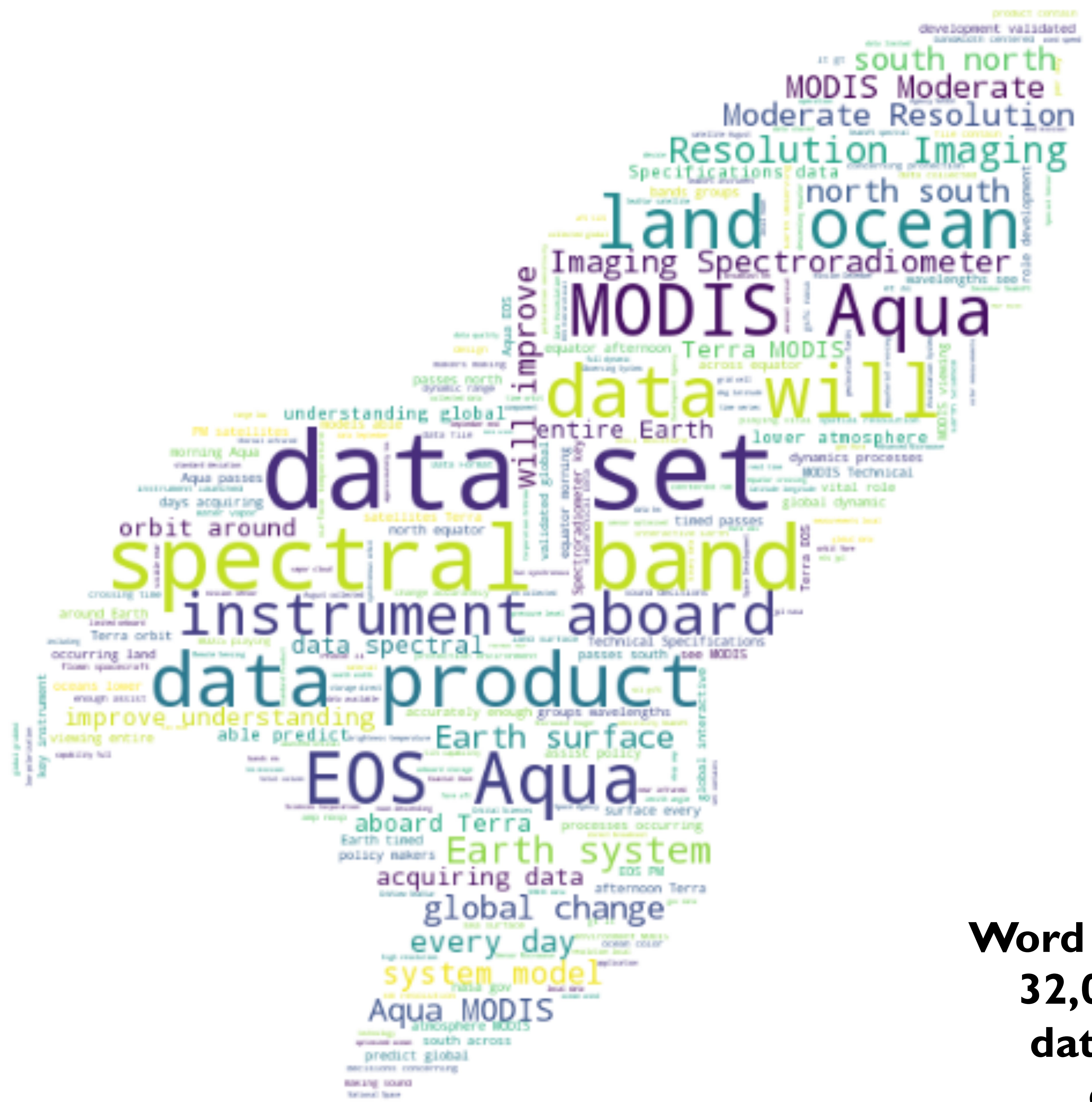
Are Descriptions, Keywords representative of the content?

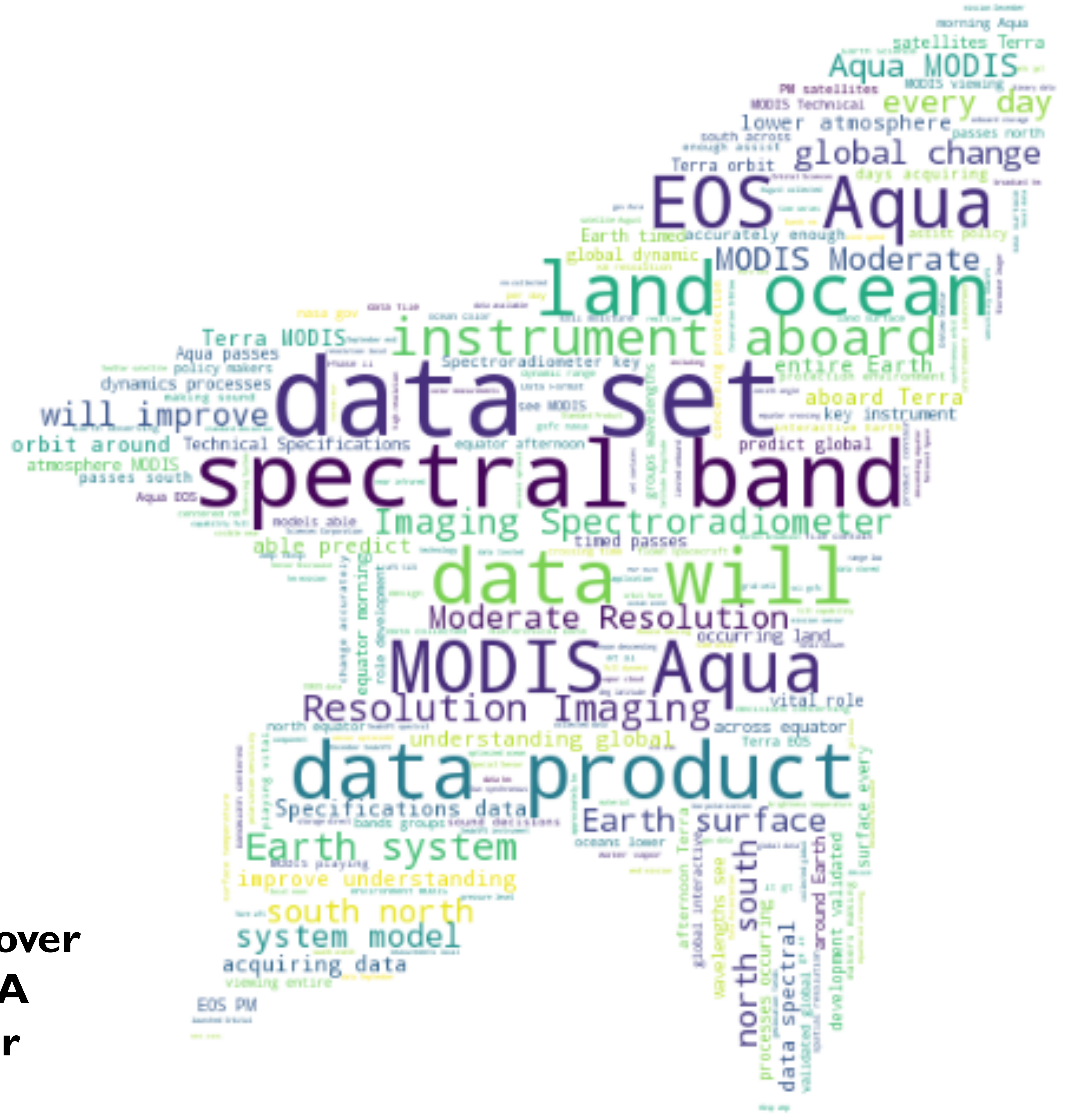
Natural Language Processing (NLP) Python Libraries

- NLTK
- TextBlob
- spaCy
- gensim
- Stanford CoreNLP

Jupyter Notebooks:
https://github.com/nderzsy/NASADatanauts

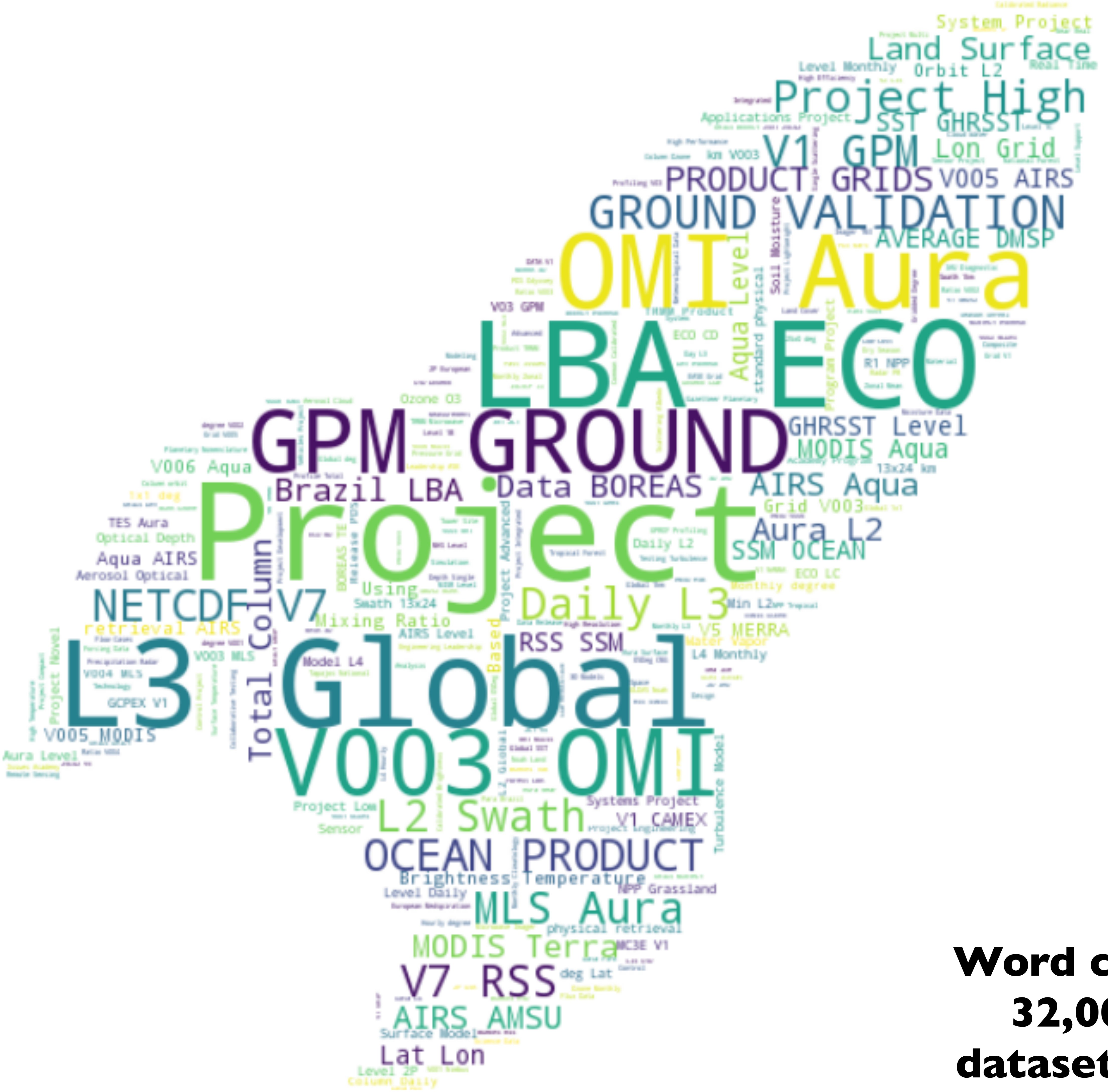
Word Clouds in Description





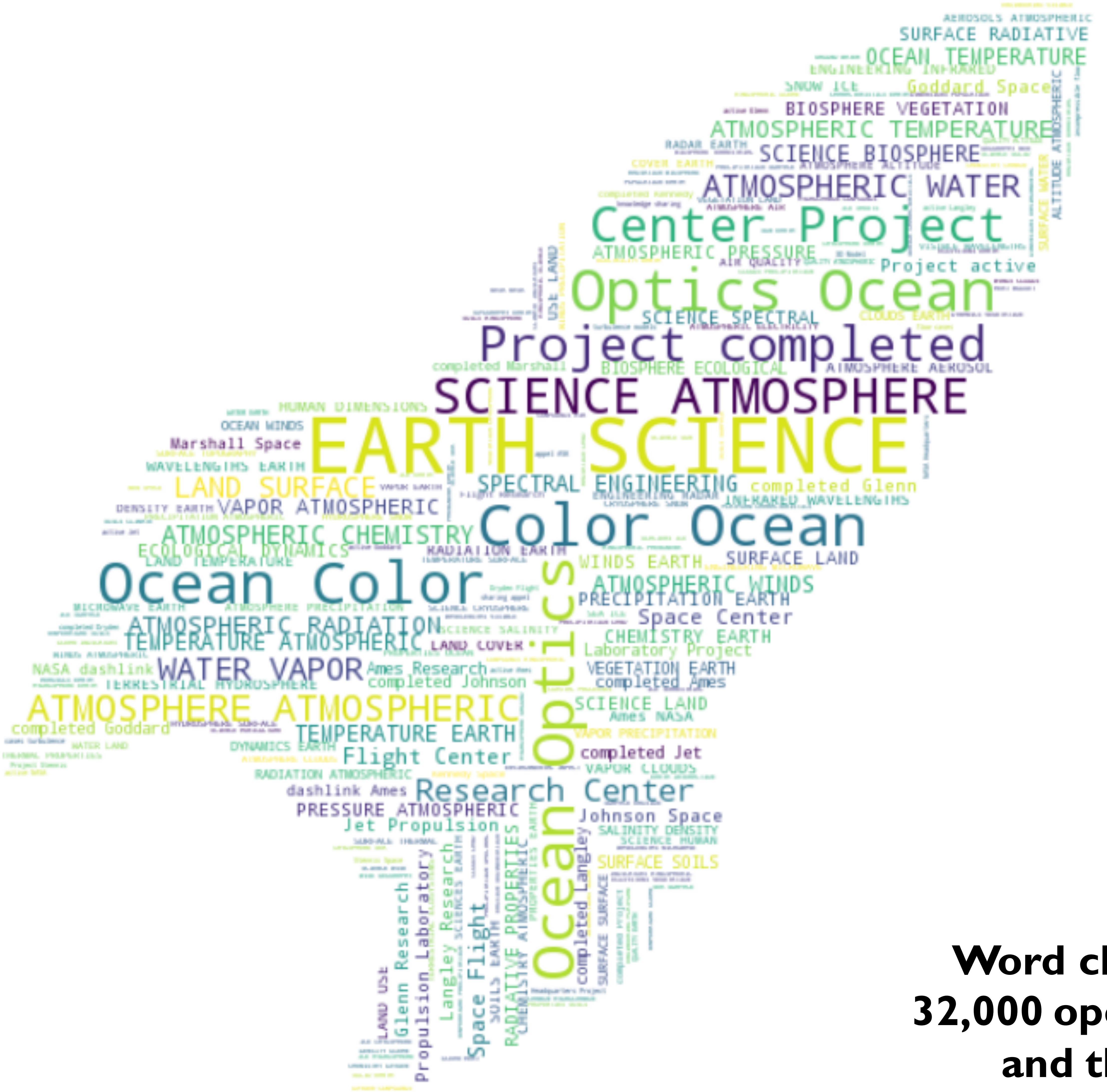
Word clouds of the over 32,000 open NASA datasets and their descriptions

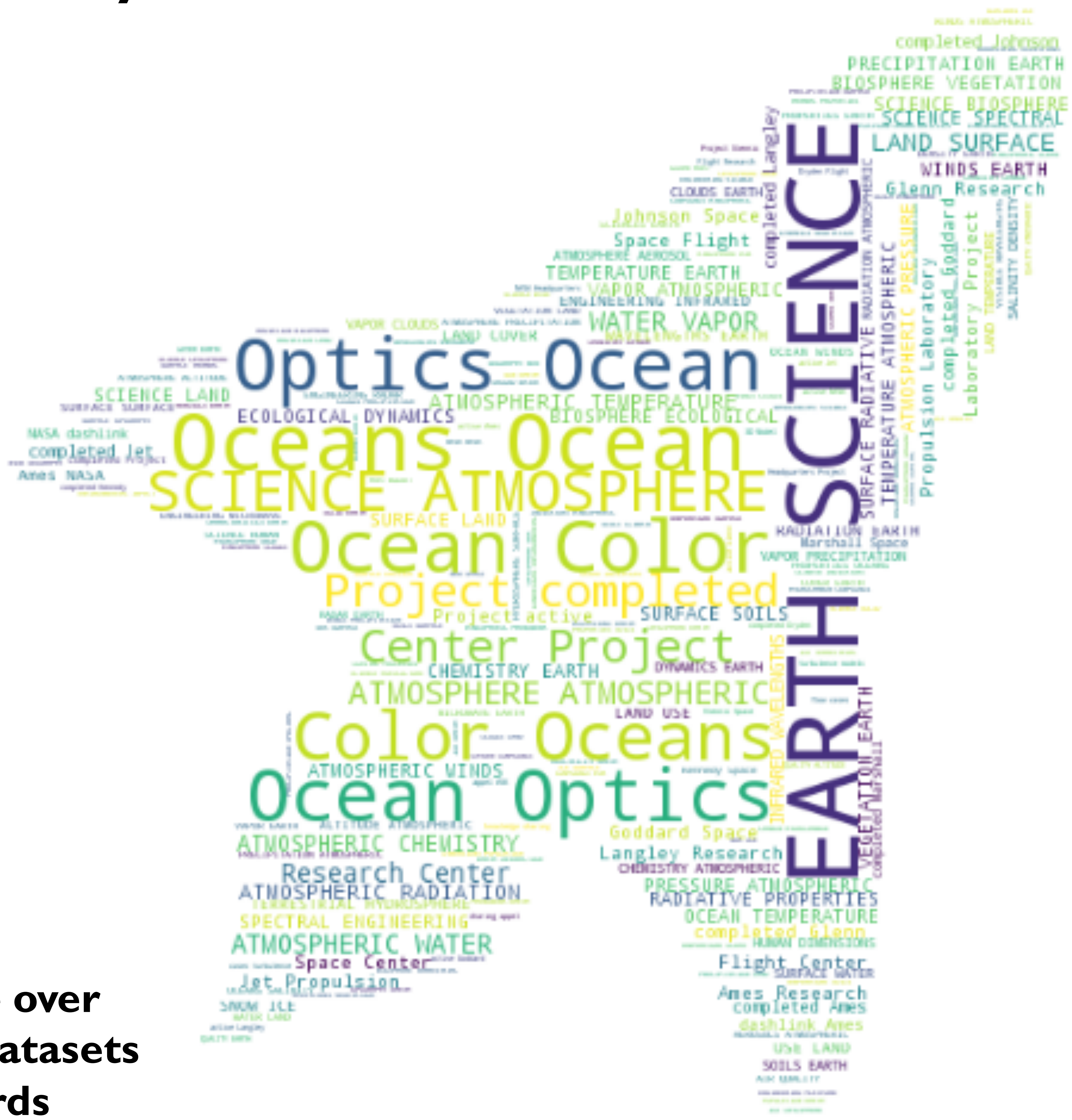
Word Clouds in Title



Word clouds of the over 32,000 open NASA datasets and their titles

Word Clouds in Keywords

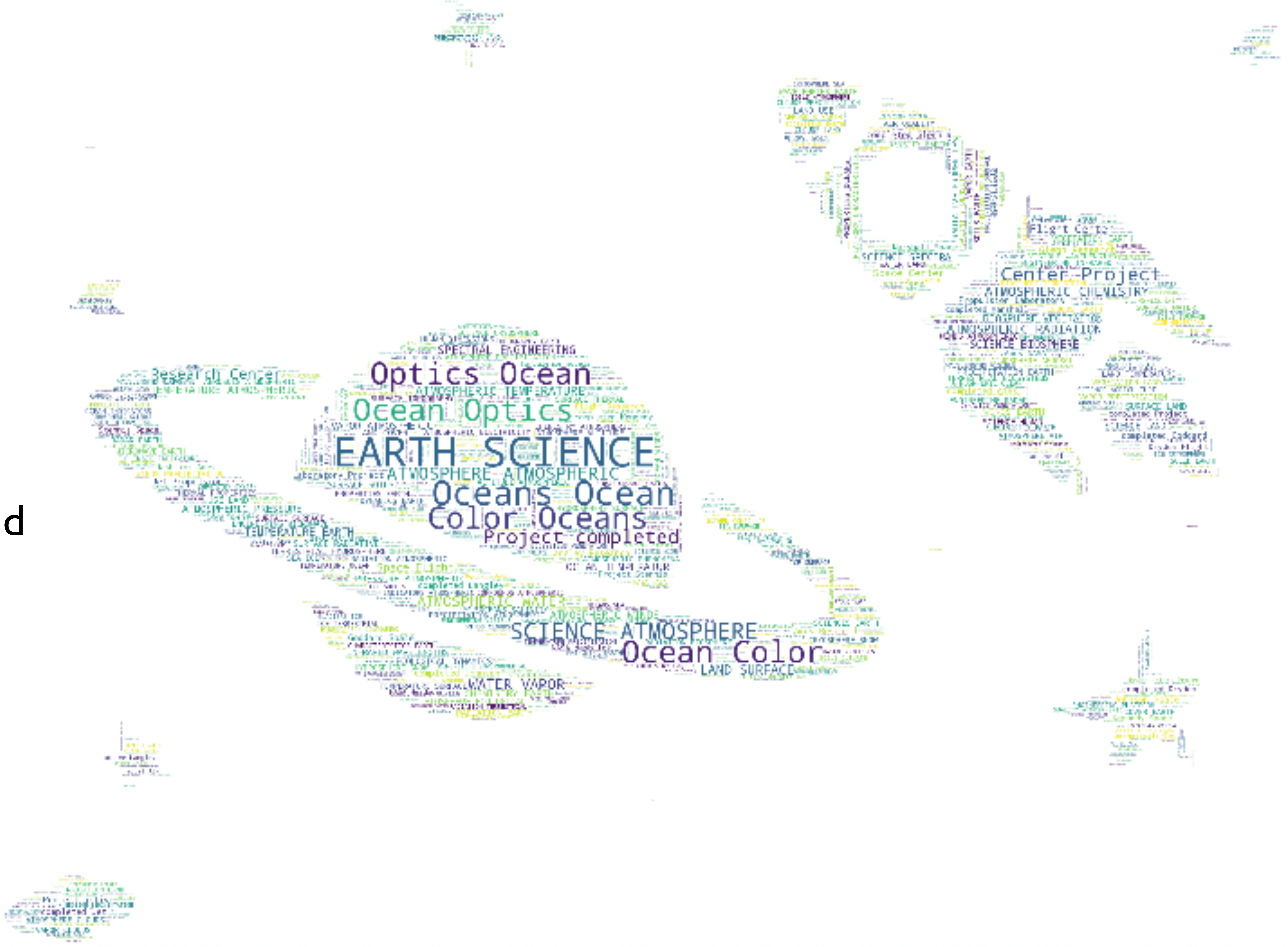




Word clouds of the over 32,000 open NASA datasets and their keywords

How to Obtain Customized Word Clouds?

- get stencil (shape of your choice)
- get text
- https://github.com/amueller/word_cloud



Text Preprocessing, Cleaning

- treat “Data” and “data” as identical words: covert all words to lowercase lower()
- remove special characters, codes, numbers: regular expressions
- check for misspelling
- stop words: this, and, for, where, etc.
- “system” vs. “systems”: lemmatize
- “compute”, “computer”, “computation” -> “comput”: stem
- tokenize: break down text to smallest parts (words)
- POS tagging
- dimensionality reduction (PCA)

Stemming

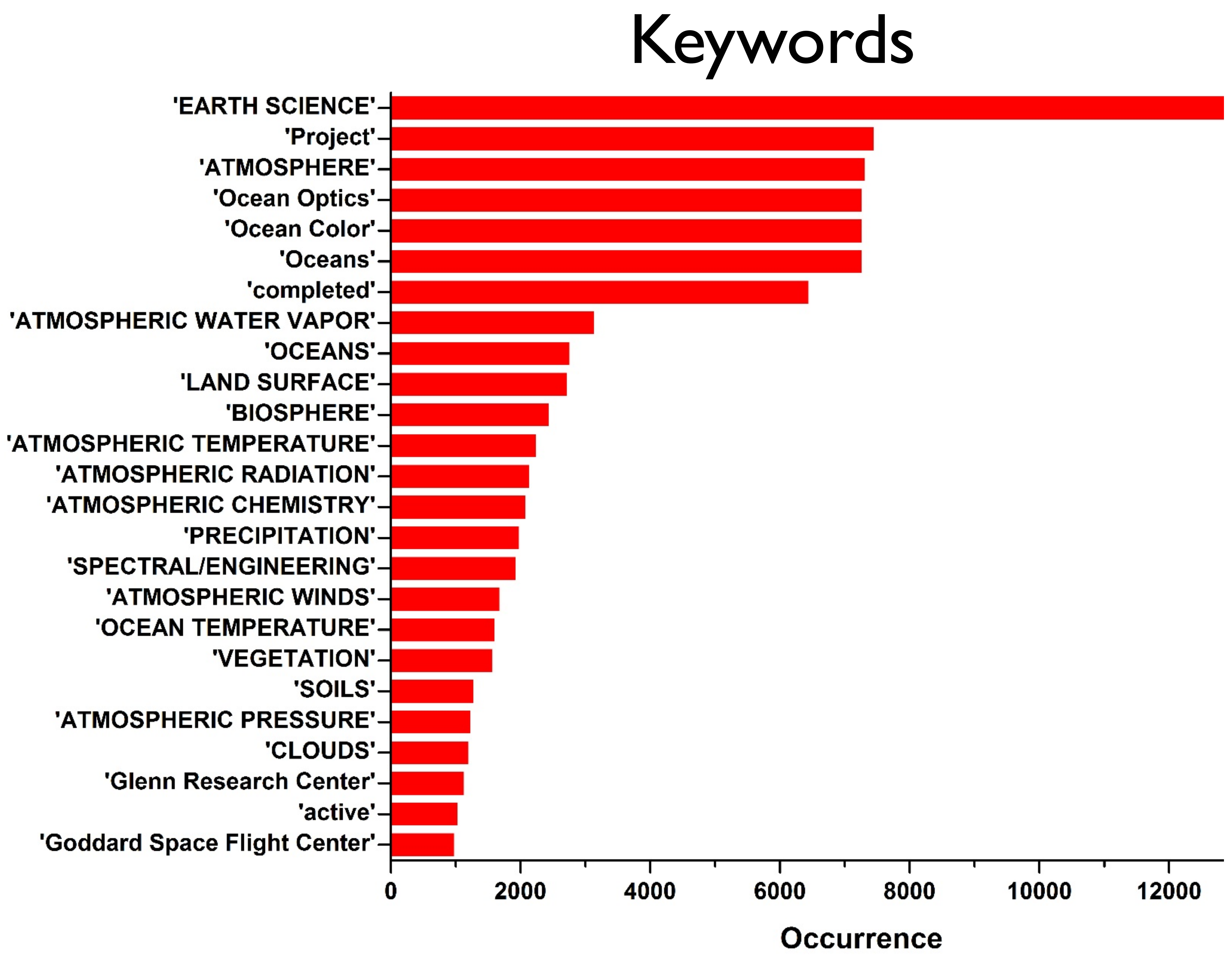
- **from** nltk.stem.api **import** StemmerI
- **from** nltk.stem.regexp **import** RegexpStemmer
- **from** nltk.stem.lancaster **import** LancasterStemmer
- **from** nltk.stem.isri **import** ISRIStemmer
- **from** nltk.stem.porter **import** PorterStemmer
- **from** nltk.stem.snowball **import** SnowballStemmer
- **from** nltk.stem.wordnet **import** WordNetLemmatizer
- **from** nltk.stem.rslp **import** RSLPStemmer

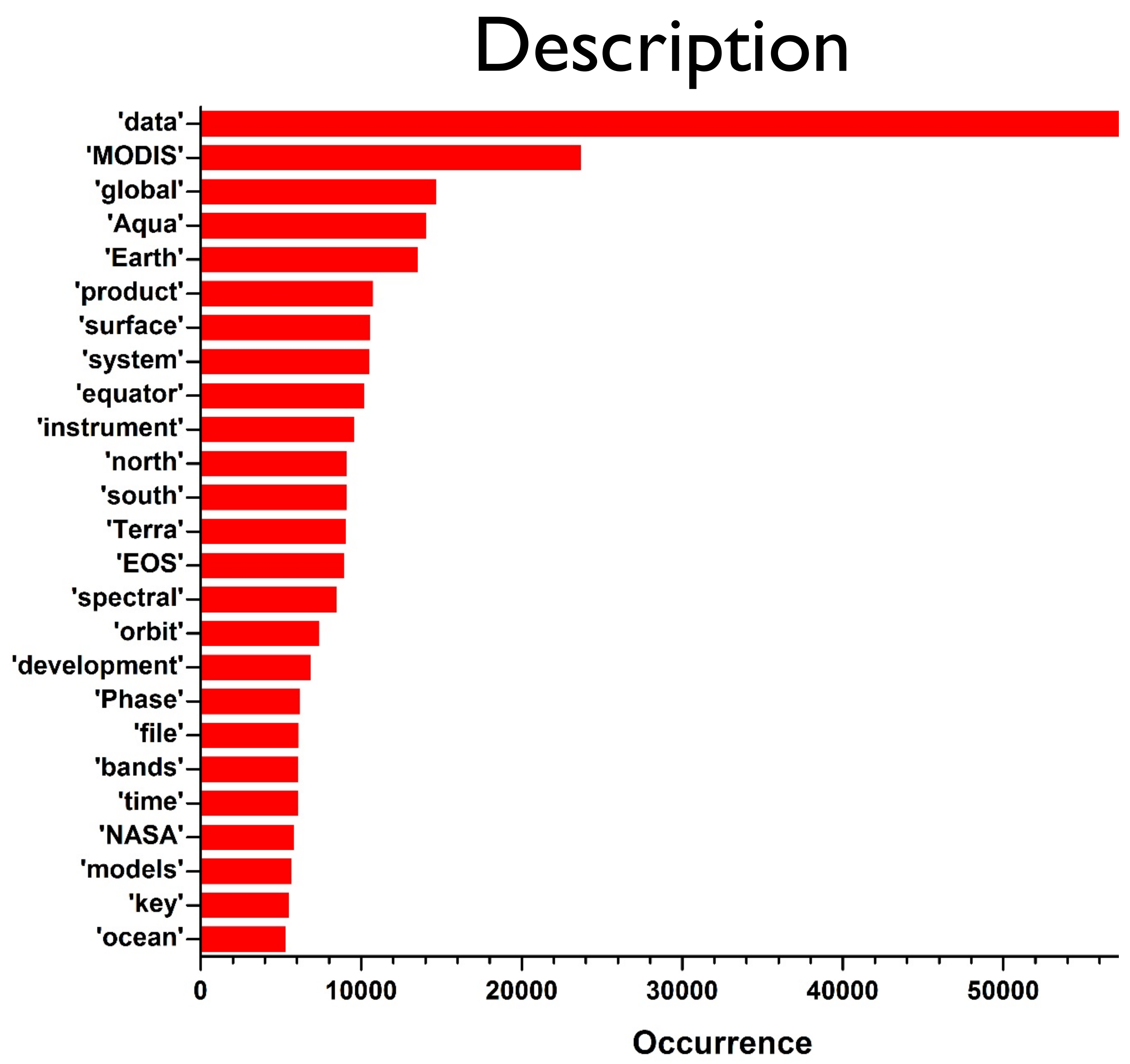
Lemmatization

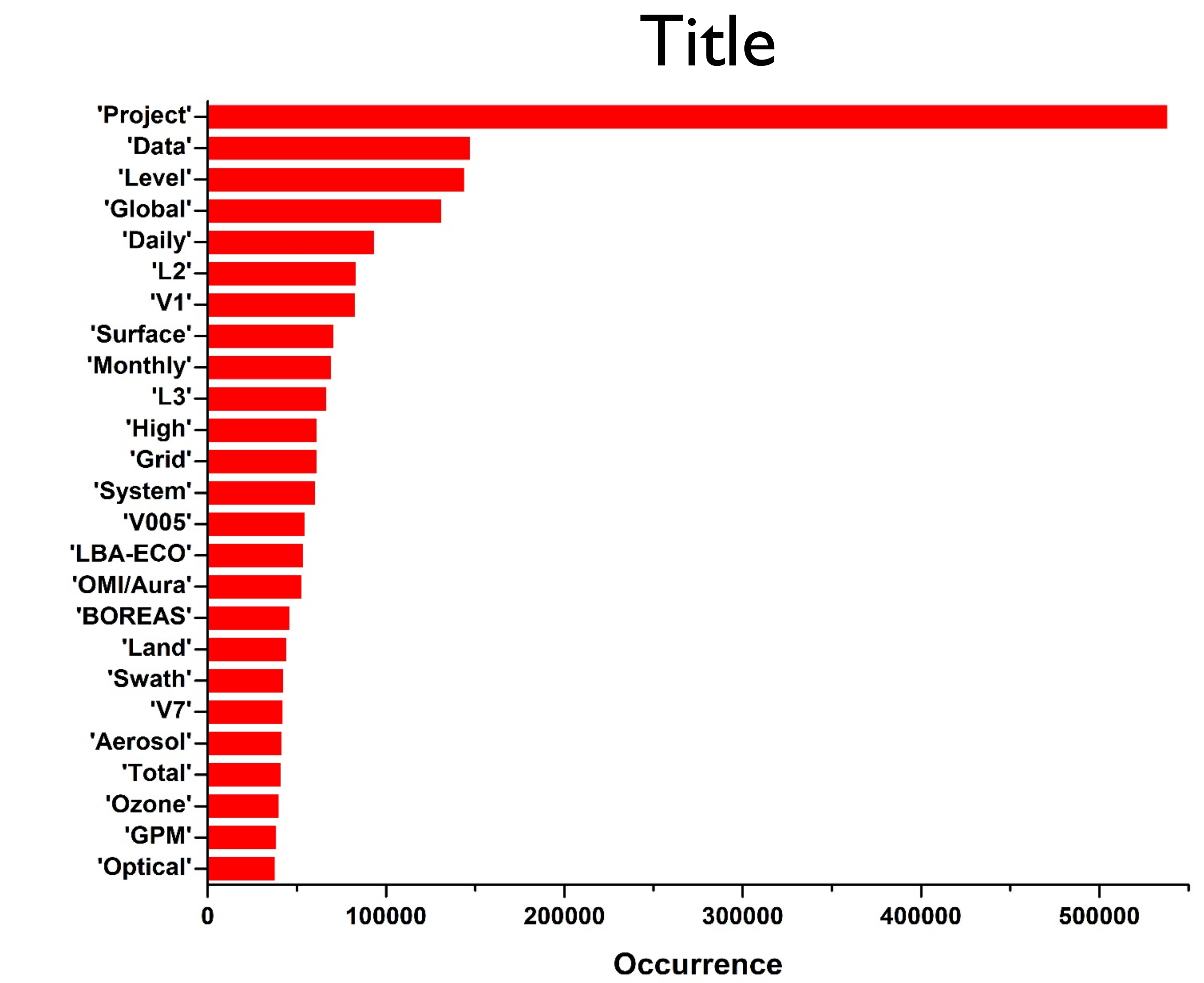
- Lemmatization: similar to stemming, but with stems being valid words
- nltk.WordNetLemmatizer()

Term Frequency

- the number of times a word occurs in text corpus

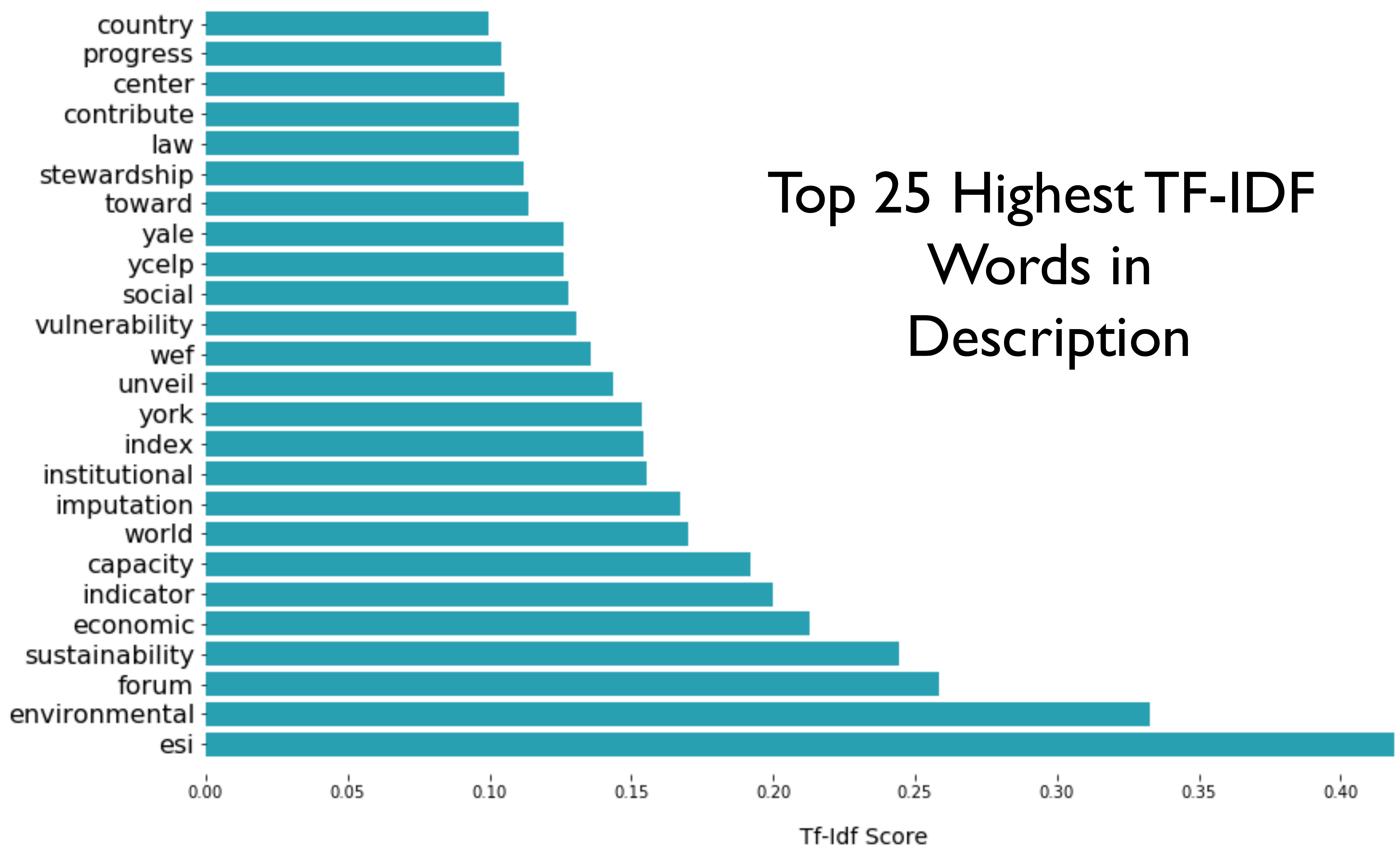




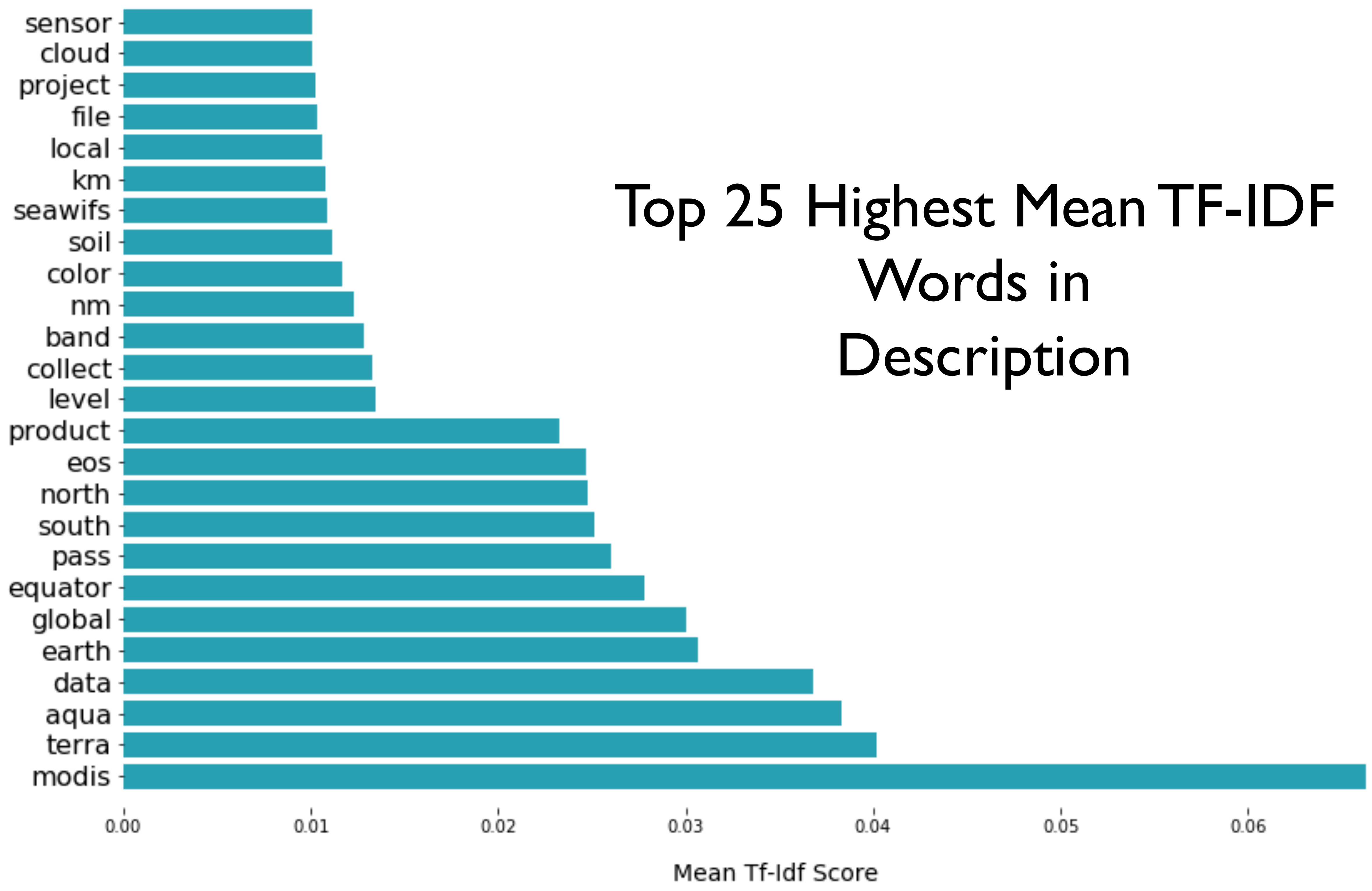


TF-IDF

- term frequency – inverse document frequency
- measures term frequency / document frequency

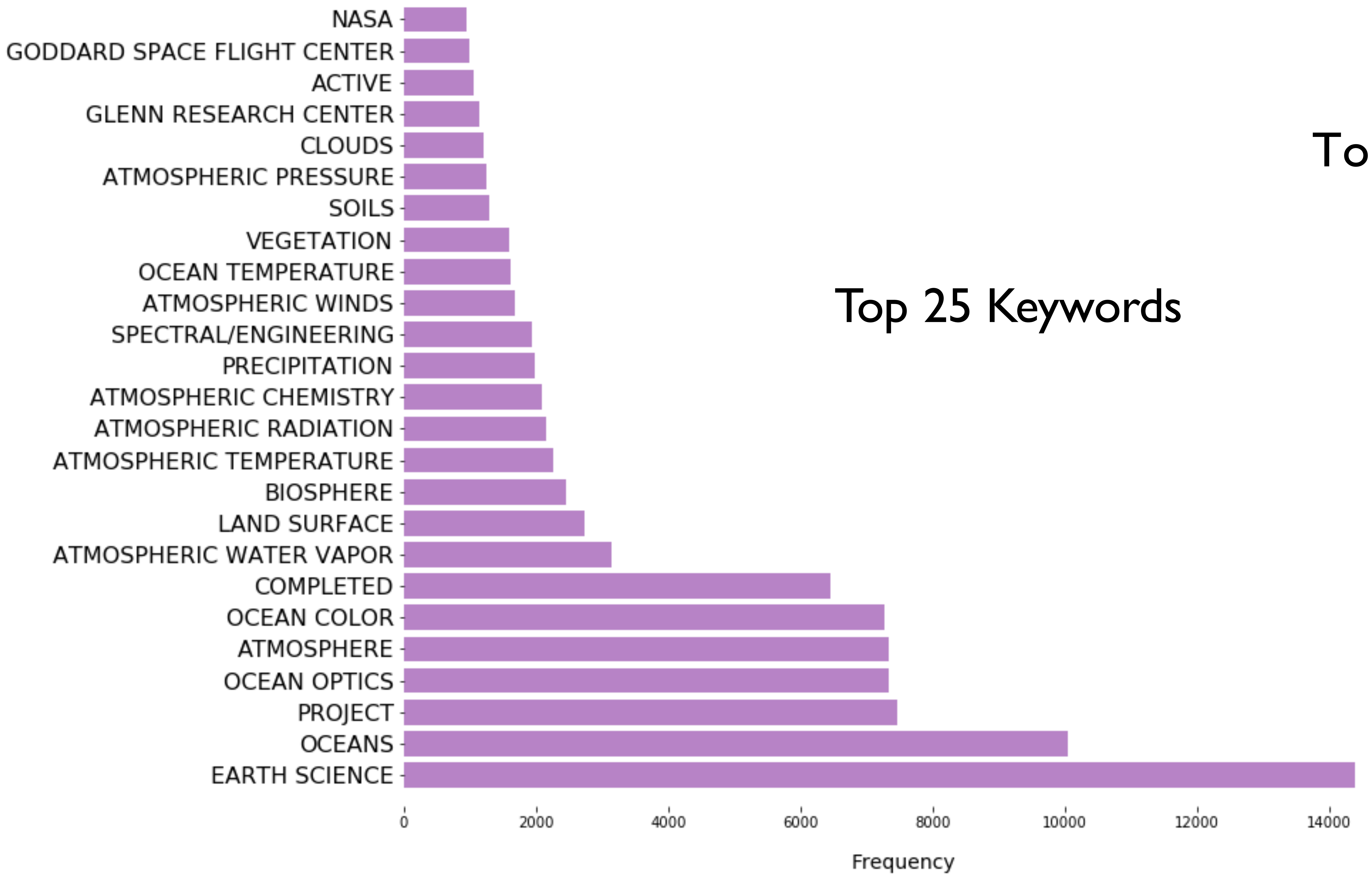


Top 25 Highest TF-IDF Words in Description



Top 25 Highest Mean TF-IDF Words in Description

Description TF-IDF and Keywords



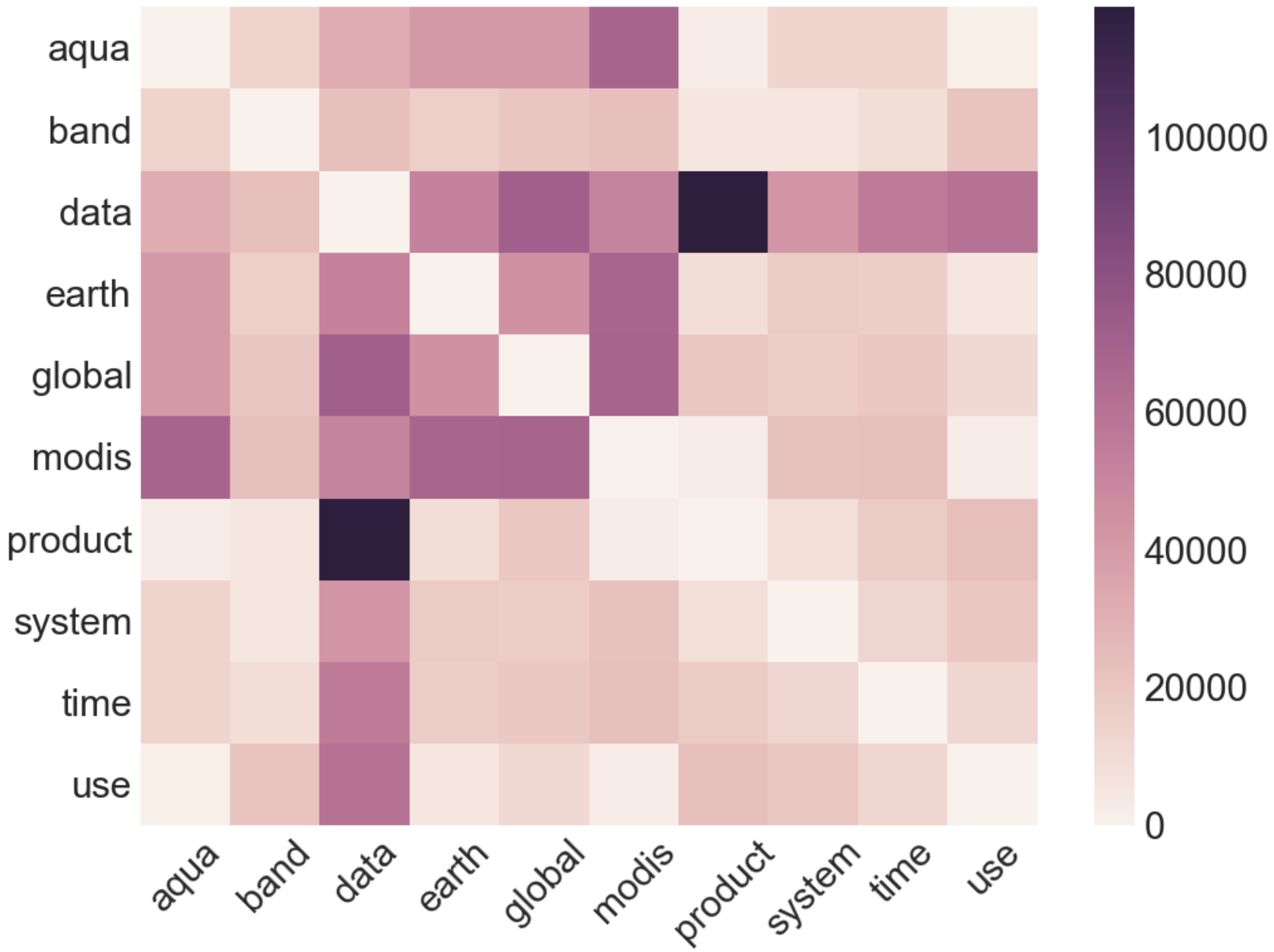
Top 10 keywords:

1. EARTH SCIENCE: 14387
2. OCEANS: 10034
3. PROJECT: 7464
4. OCEAN OPTICS: 7325
5. ATMOSPHERE: 7324
6. OCEAN COLOR: 7271
7. COMPLETED: 6453
8. ATMOSPHERIC WATER VAPOR: 3143
9. LAND SURFACE: 2721
10. BIOSPHERE: 2450

Word Co-Occurrence

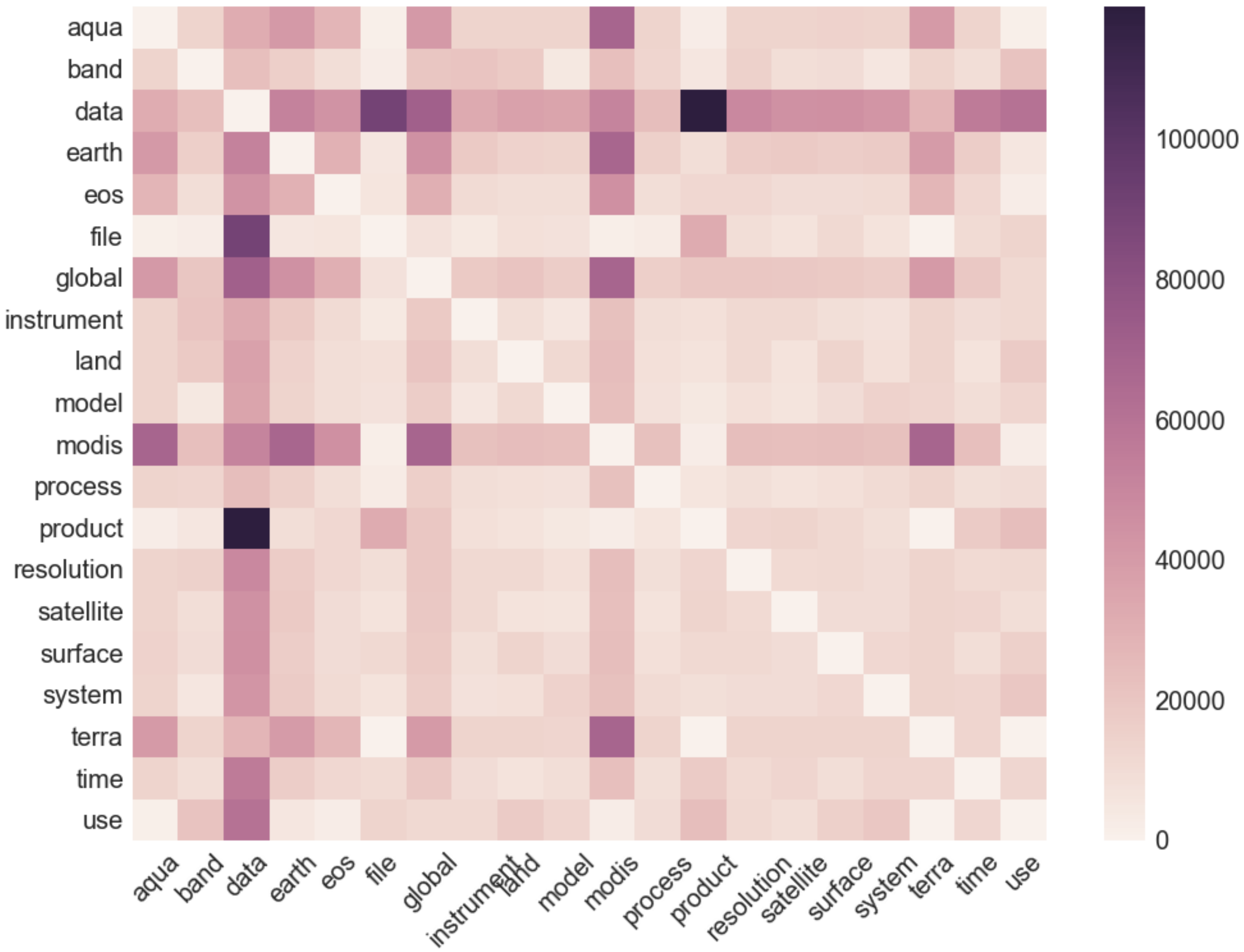
- Co-occurrence matrix of top most frequently co-occurring terms

Top 10
Word Co-Occurrence
Matrix

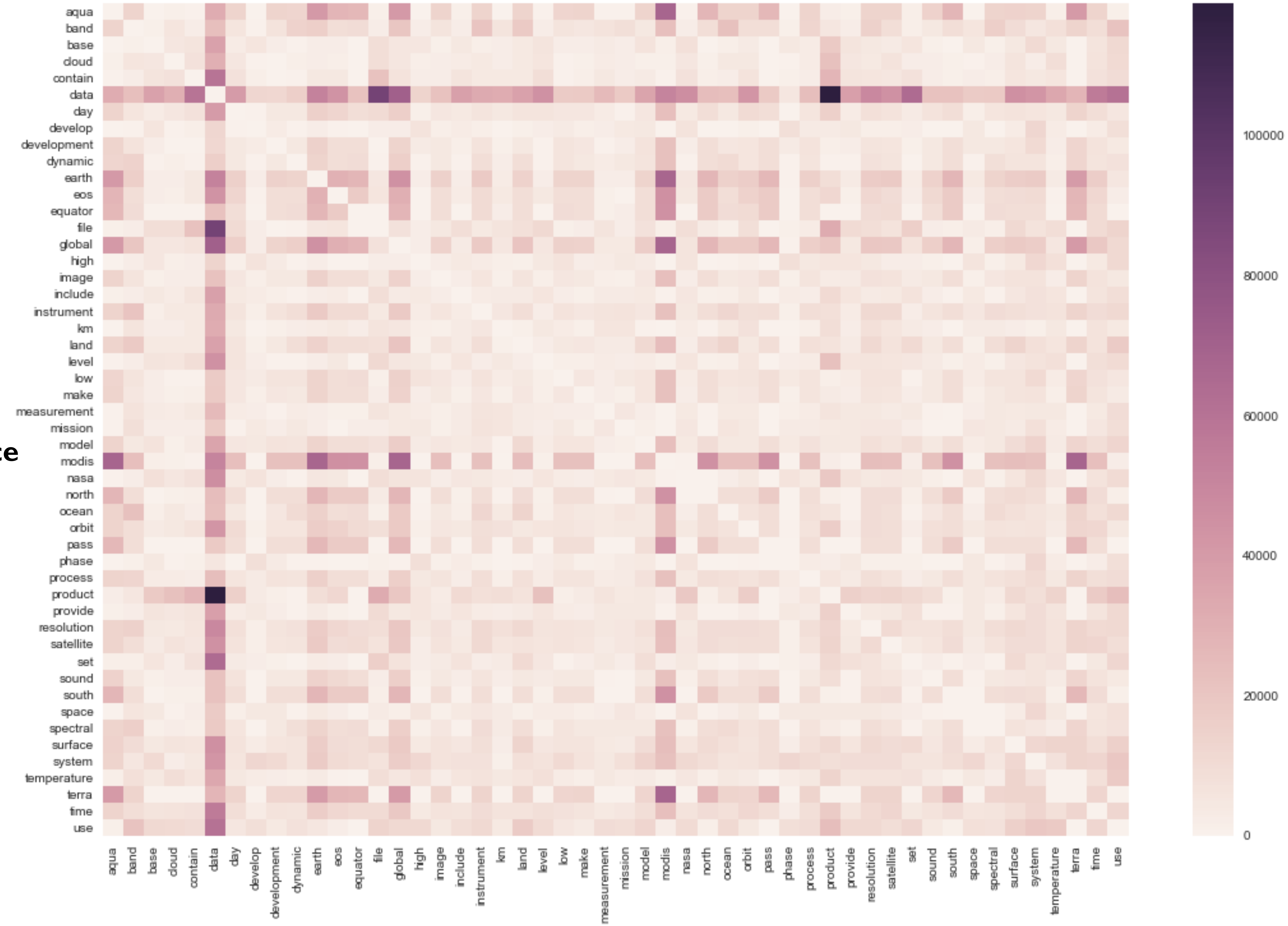


Word Co-Occurrence

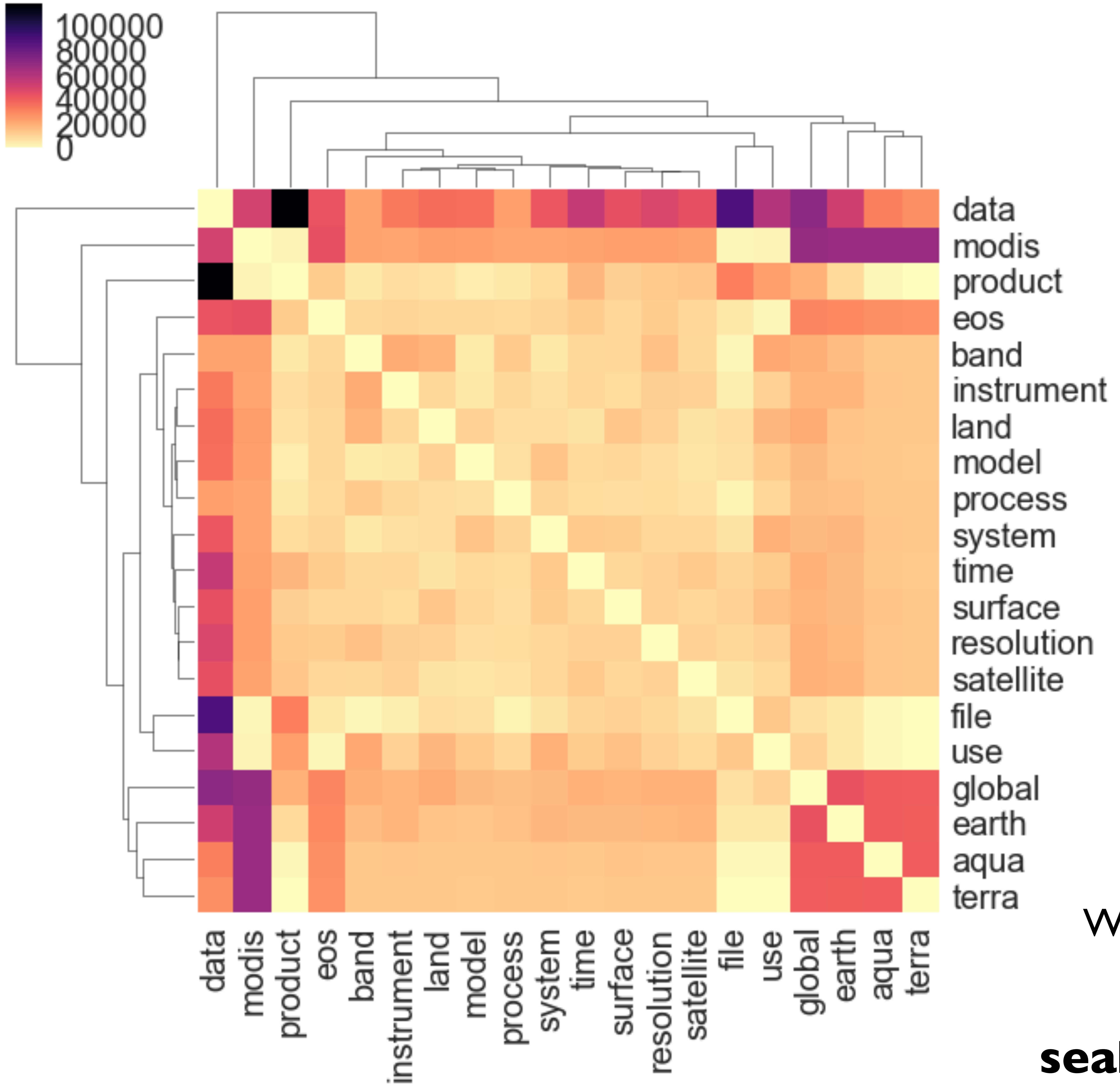
Top 20
Word Co-Occurrence
Matrix



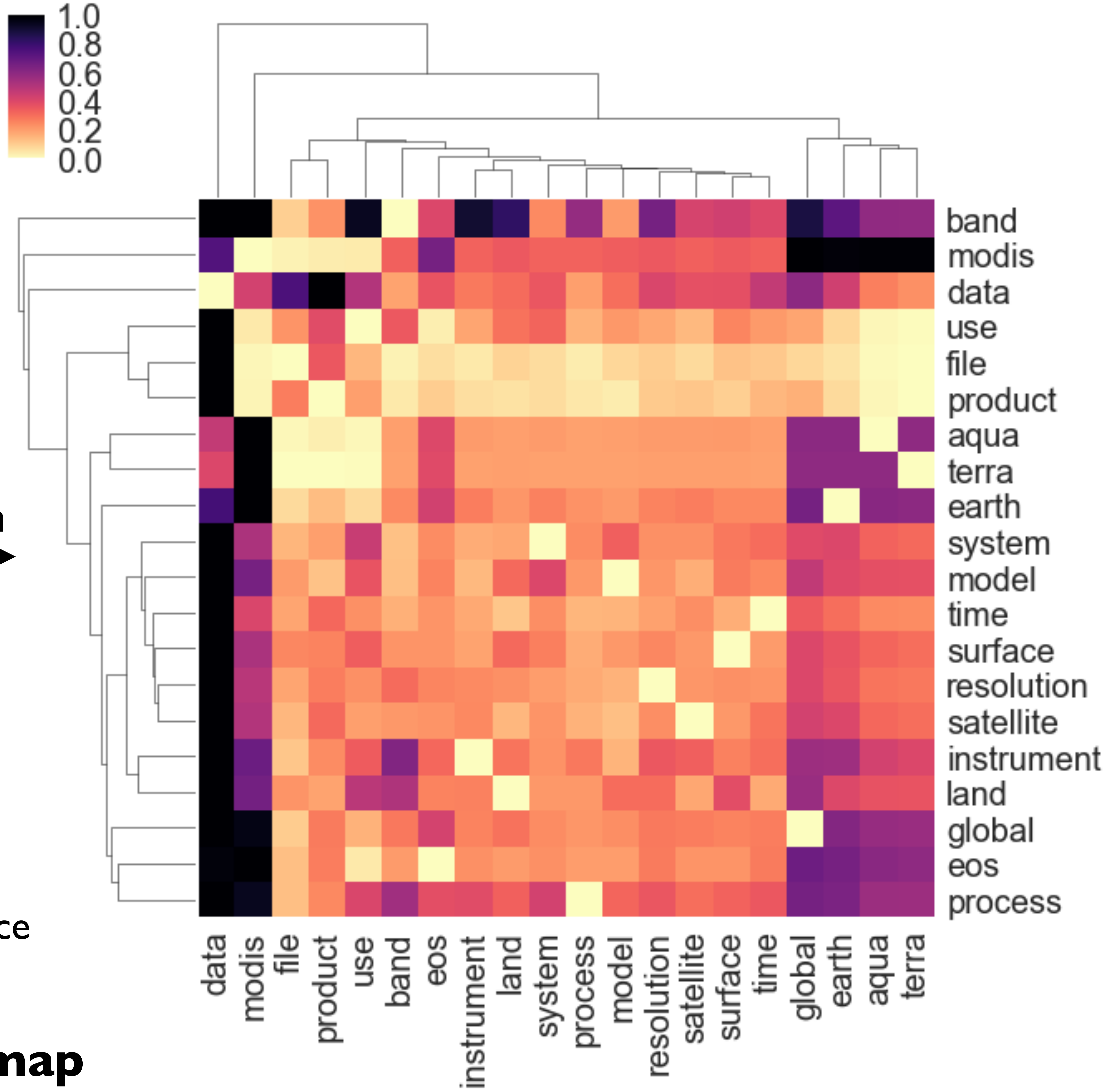
Top 50 Word Co-Occurrence Matrix



Discovering Structure in Heatmap Data



standardization



Top 20
Word Co-Occurrence
Matrix

seaborn.clustermap

Are features pulled from text (such as title, description fields)

and/or human supplied-keywords

descriptive of the content?

Topic Modeling…

What is Topic Modeling?

An efficient way to make sense of large volume of texts.

Identify topics within text corpus.

Categorize documents into topics.

Associate words with topics.

Who uses it?

Search engines, for marketing purpose, etc.

Latent Dirichlet Allocation (LDA)

- several techniques, but LDA is the most common
- Bayesian inference model that associates each document with a probability distribution over topics
- topics are probability distributions over words (probability of the word being generated from that topic for that document)
- clusters words into topics
- clusters documents into mixture of topics
- scales well with growing corpus
- before running LDA algorithm, we have to specify the number of topics: how to choose beforehand the optimal number of topics?

Topic Model Evaluation: Topic Coherence

Q: How to select the top topics?

A: Calculate the UMass topic coherence for each topic. Algorithm from *Mimno, Wallach, Talley, Leenders, McCallum: Optimizing Semantic Coherence in Topic Models, CEMNLP 2011*.

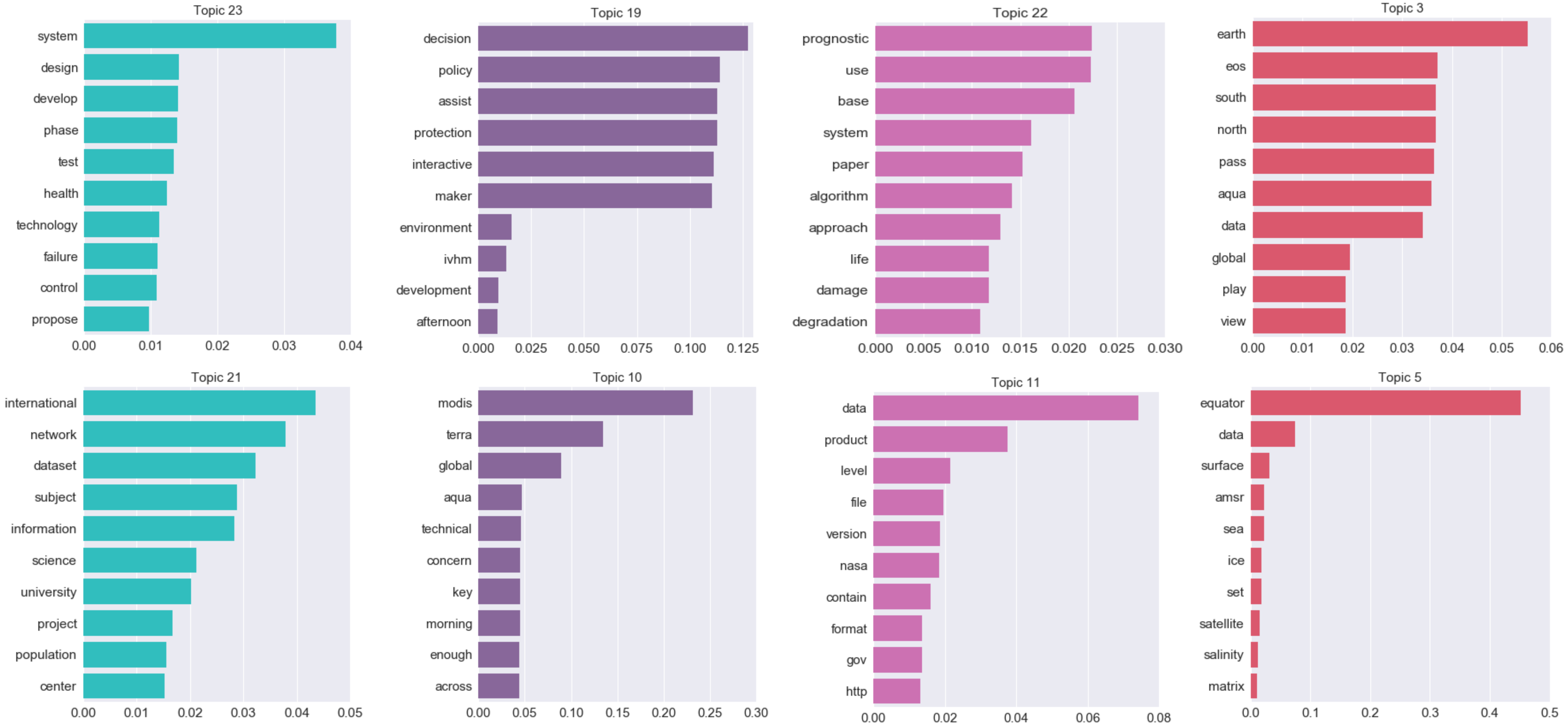
$$\text{Coherence} = \sum_{i<j} \text{score}(w_i, w_j)$$

pairwise scores on the words used to describe the topic.

$$score_{UMass(w_i,w_j)} = \log \frac{D(w_i, w_j) + 1}{D(w_i)}$$

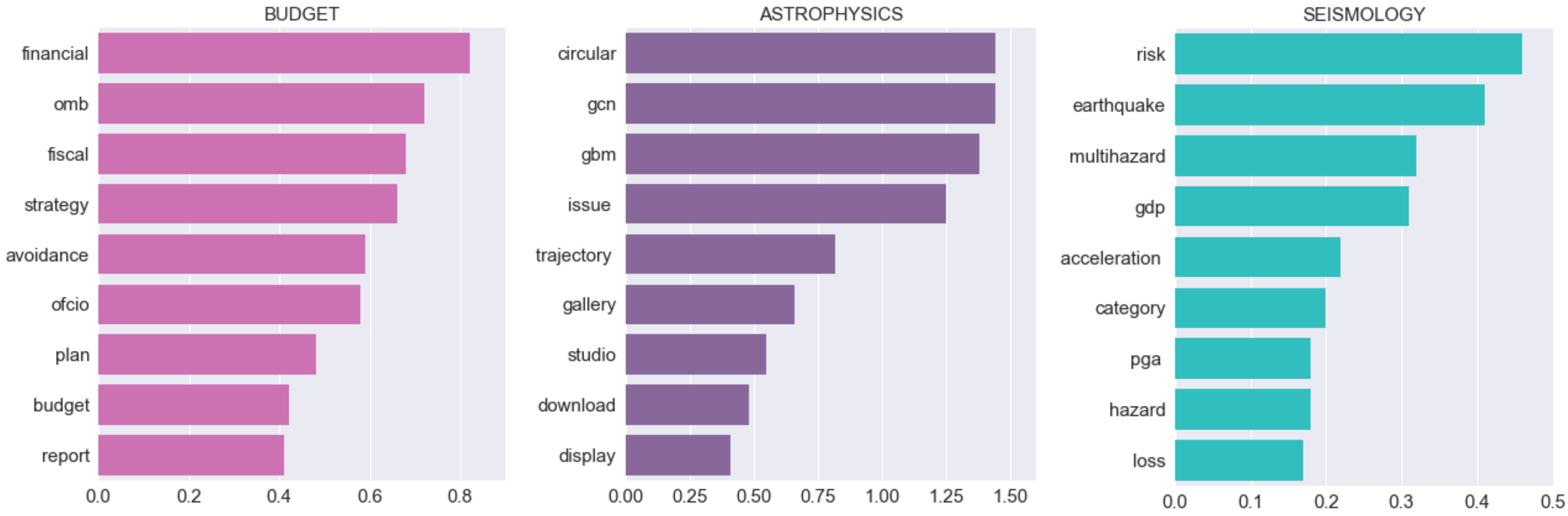
$D(w_i)D(w_i)$ as the count of documents containing the word w_iw_i, $D(w_i,w_j)D(w_i,w_j)$ the count of documents containing both words w_iw_i and w_jw_j, and DD the total number or documents in the corpus.

openNASA Topics of Highest Coherence



Keywords for Topics

- selected keywords with their most frequently occurring terms



Other Clustering Method: K-Means

- using TF-IDF, the document vectors are put through a K-Means clustering algorithm which computes the Euclidean distances amongst these documents and clusters nearby documents together
- the algorithm generates cluster tags, known as cluster centers which represent the documents within these clusters
- K-means distance:
  - Euclidean
  - Cosine
  - Fuzzy
- Accuracy comparison:
- silhouette analysis can be used to study the separation distance between the resulting clusters; can be used to determine the optimal number of clusters

K-Means Clustering

- Top 10 terms per cluster:

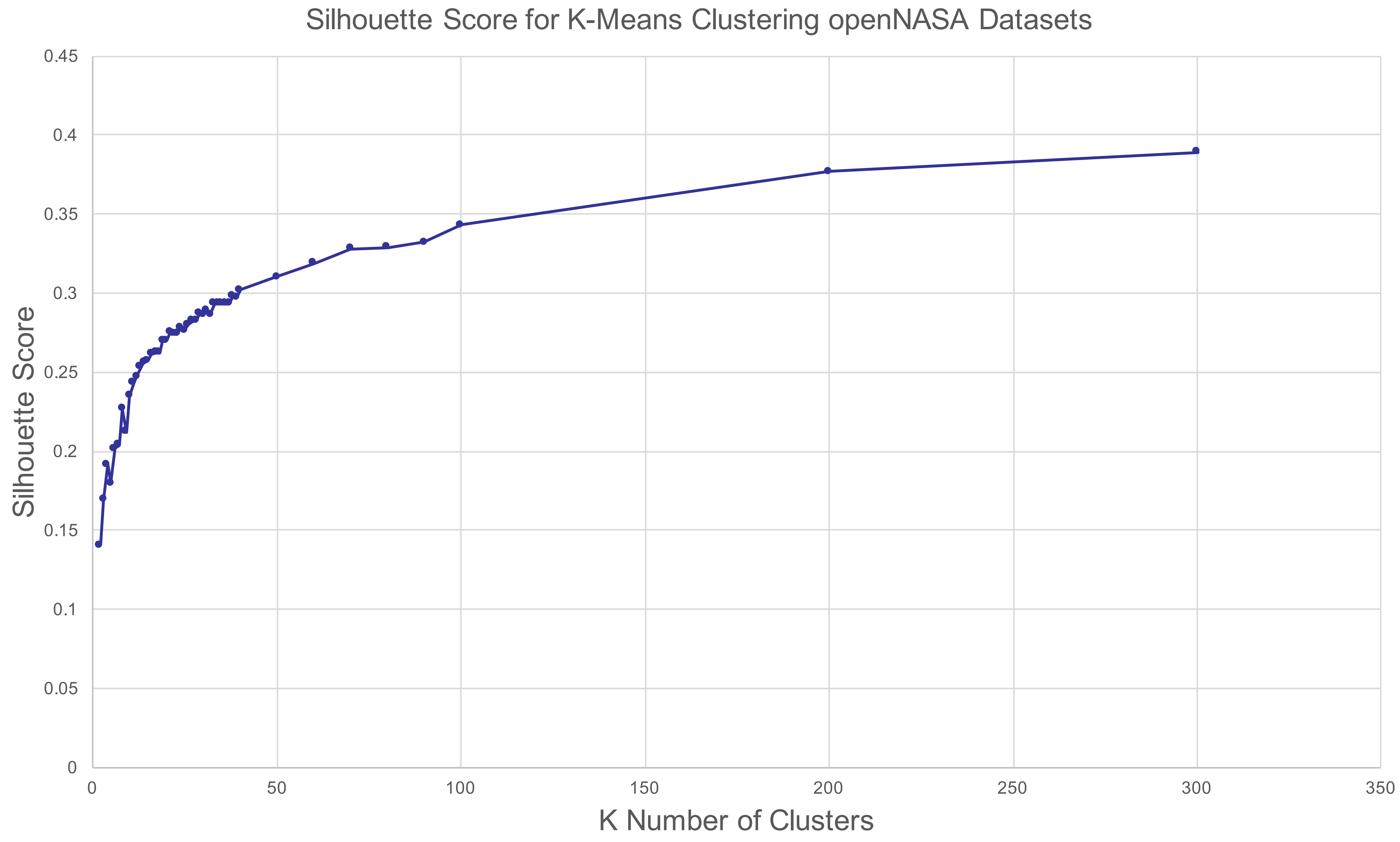
Cluster 0	Cluster 1	Cluster 2	Cluster 3	Cluster 4
seawifs	data	modis	band	product
local	project	terra	oct	data
collect	system	aqua	color	level
km	use	earth	adeos	version
mission	soil	global	czcs	file
data	high	pass	nominal	aquarius
orbview	contain	south	sense	set
seastar	phase	north	spacecraft	daily
broadcast	set	equator	agency	ml
noon	gt	eos	thermal	standard

- Similar clusters with top words as found using LDA

K-Means Clustering

- cluster size distribution for k = 5

Cluster 2:	19031
Cluster 0:	5805
Cluster 1:	4481
Cluster 3:	1968
Cluster 4:	804

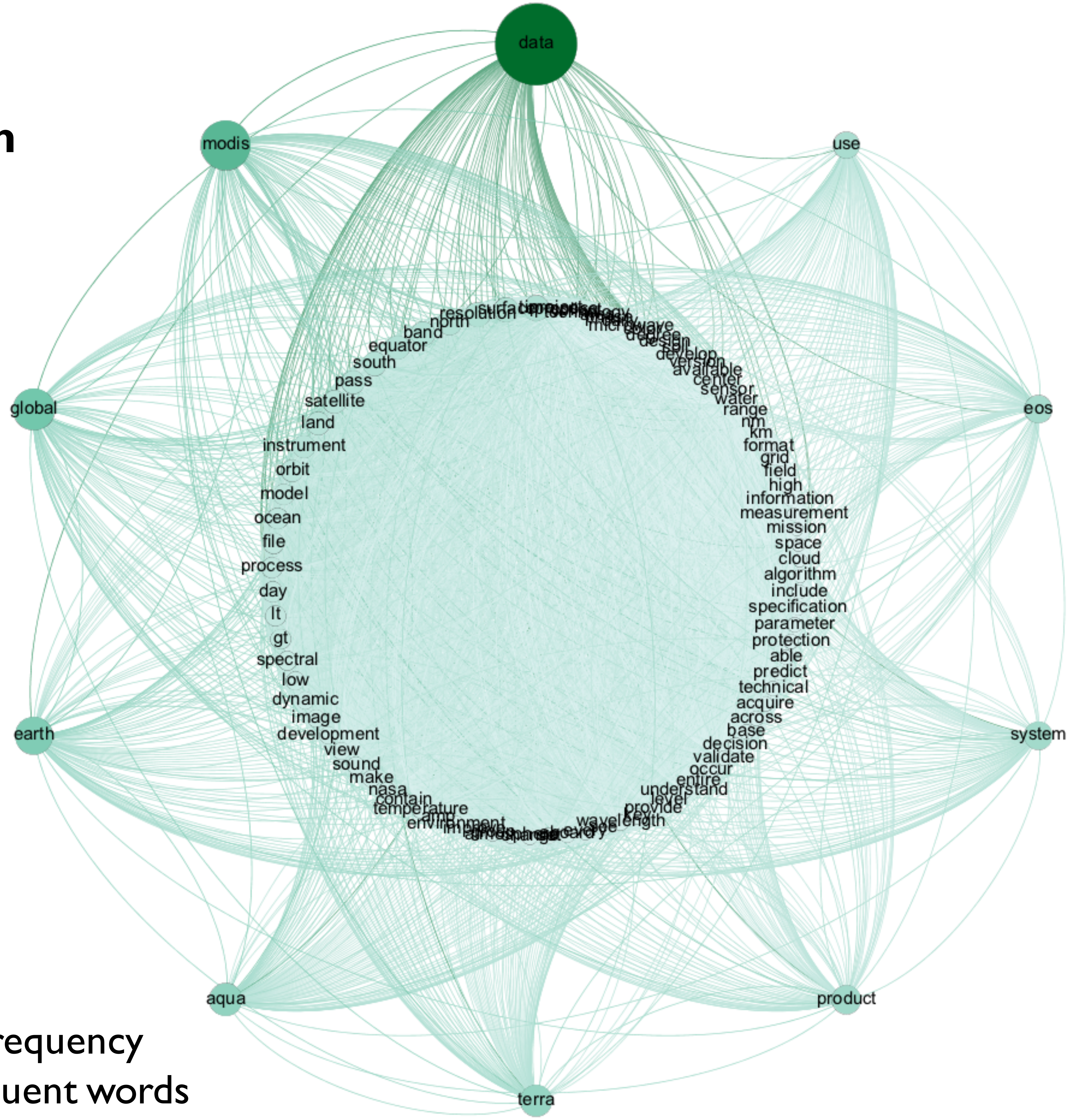


- silhouette score improves by increasing the number of clusters, best performance for K > 100
- analysis of cluster size distributions with varying K can reveal additional information

Text Classification

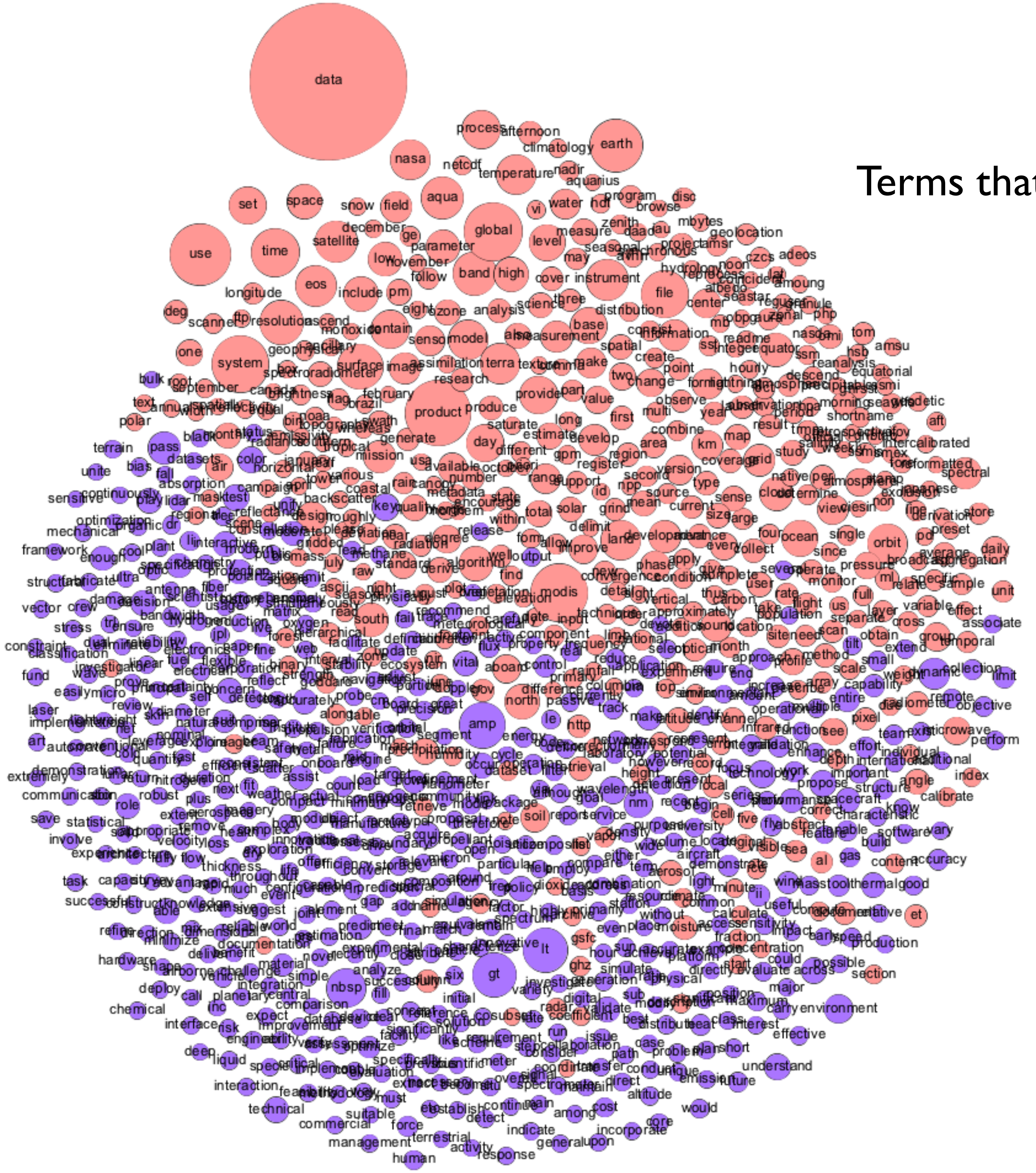
- Naïve Bayes probabilistic model
  - Multinomial model
    - data follows multinomial distribution
    - each feature value is a count (word co-occurrence, weight, tf-idf, etc.)
    - Take into account word importance
  - Bernoulli model
    - data follows a multivariate Bernoulli distribution
    - bag of words (count word occurrence)
    - each feature is a binary feature (word in text? True/False)
    - ignores word significance
- KNN (K-nearest neighbor) classifier
- Decision trees
- Support Vector Machine (SVM)

Top 100
most frequent words in Description



Nodes with the highest connectivity are not all the same as most frequently co-occurring

- size/color hue ~ occurrence frequency
- outer circle: top 10 most frequent words



Terms that occur > 1% of the documents (in more than 320 data set descriptions)

- Network size: # of nodes (words) = 1015
- Average degree: 890.319 (densely connected terms)
- Components: 1 connected component
- Modularity analysis: 2 clusters (pink, violet)

Bipartite Network

- Term in Description – Keyword:
  - connect 2 words in description if they appear under same Keyword
  - connect 2 keywords if they share common words in Description
- All bipartite graph projection possibilities:
  - description – keyword
  - description – title
  - title – keyword
- Type of interaction:
  - link direction: directed/undirected
  - link weight: how strong the connection (co-occurrence matrix)

Structural Network Properties

- degree distribution: reveals connectivity pattern
- do we have hubs (nodes with significantly higher number of connections)?
- average degree (how interconnected the elements are)
- clusters, community detection (subgroups with elements more densely connected among each other)

Network structures:

powerful tool for

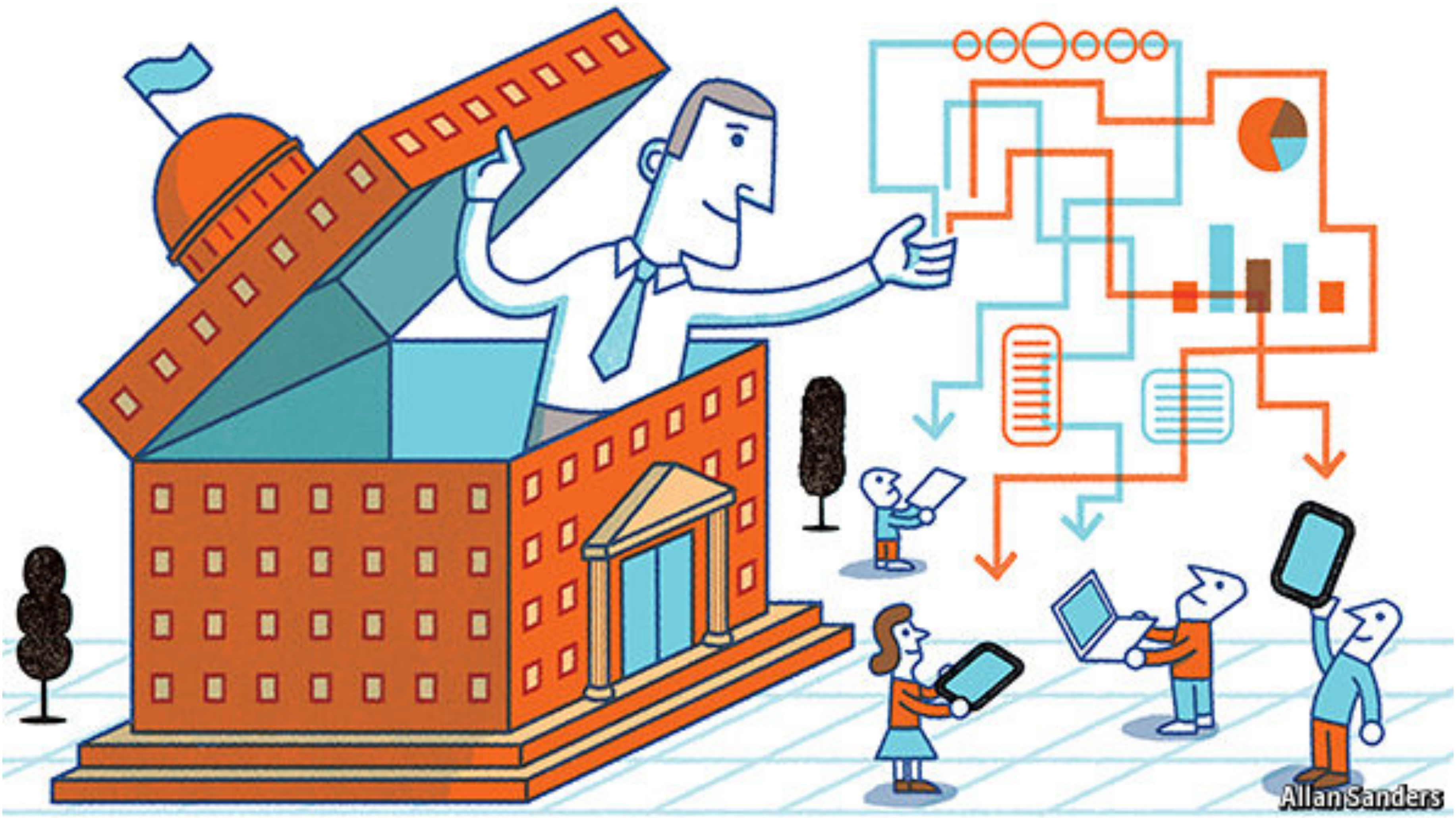
capturing additional information about

interconnected systems!

What are the important connections between
NASA datasets and other important datasets outside of NASA
in the US government?

Other Open Government Dataset Collections

http://data.nasa.gov/data.json
http://www.epa.gov/data.json
http://data.gov
http://www.nsf.gov/data.json
http://usda.gov/data.json
http://data.noaa.gov/data.json
http://www.commerce.gov/data.json
http://nist.gov/data.json
http://www.defense.gov/data.json
http://www2.ed.gov/data.json
http://www.dol.gov/data.json
http://www.state.gov/data.json
http://www.dot.gov/data.json
http://www.energy.gov/data.json
http://nrel.gov/data.json
http://healthdata.gov/data.json
http://www.hud.gov/data.json
http://www.doi.gov/data.json
http://www.justice.gov/data.json



Open Government Data: Out of the Box
The Economist

http://www.archives.gov/data.json
http://www.nrc.gov/data.json
http://www.nsf.gov/data.json
http://www.opm.gov/data.json
https://www.sba.gov/sites/default/files/data.json
http://www.ssa.gov/data.json
http://www.consumerfinance.gov/data.json
http://www.fhfa.gov/data.json
http://www.imls.gov/data.json
http://data.mcc.gov/raw/index.json
http://www.nitrd.gov/data.json
http://www.ntsb.gov/data.json
http://www.sec.gov/data.json
https://open.whitehouse.gov/data.json
http://treasury.gov/data.json
http://www.usaid.gov/data.json
http://www.gsa.gov/data.json

https://www.economist.com/news/international/21678833-open-data-revolution-has-not-lived-up-expectations-it-only-getting

pyNASA and pyOpenGov Libraries

- Python library that loads all the open NASA or other government metadata collection at once

pyNASA

https://github.com/bmtgoncalves/pyNASA

pyOpenGov

https://github.com/nderzsy/pyOpenGov

How to install:

>> pip install pyNASA

>> pip install pyOpenGov

Takeaways

- NLP enables understanding of structured and unstructured text
- Topic modeling useful tool for understanding topics in large text corpus, documents
- Topic models efficient for evaluating the accuracy of human-supplied descriptions, keywords
- Tedious preprocessing a must before modeling (stop words, lemmatization, special characters, etc.)
- Network (graph) structure can reveal more information, term/topic associations
- Network projections allow to answer additional questions
- Open government data enables citizens in understanding topics, areas of focus

Contact

GitHub: https://github.com/nderzsy/NASADatanauts

Twitter: @NoemiDerzsy

Website: http://www.noemiderzsy.com

Strata
DATA CONFERENCE