Predicting if a paper speaks about a mineral

Ramazan Bahrami

Seminar on NLP,Digital Humanities and Information Science

Tuesday 28th February, 2023

GEORG-AUGUST-UNIVERSITÄT GÖTTINGEN IN PUBLICA COMMODA SEIT 1737

Contents

Motivation

- GEOROC Papers
- Extracted Metadata from the papers

Motivation

- GEOROC Papers
- Extracted Metadata from the papers

```
In [15]: meta_data = pd.read_csv('data/2022-12-4EZ7ID_METADATA.csv',low_memory=False)
         len(meta_data)

Out[15]: 626611
```

```
In [19]: import csv

         with open("data/2022-12-4EZ7ID_CITATION.tab") as fruits_file:
             tsv_reader = csv.reader(fruits_file, delimiter="\t")

             # CITATIONS,DOI,AUTHORS,YEAR,TITLE,JOURNAL,VOL,ISSUE,PAGES,BOOK_TITLE,EDITOR,PUBLISHER,FORMATTED_CITATION)
             j=0
             next(tsv_reader)
             for  t in tsv_reader:
                 j=j+1
             print(j)

         20882
```

Motivation

- GEOROC Papers
  Reference

- Extracted Metadata from the papers
  Observation

MetaData

- Extracted Metadata from the papers
  Observation

CITATIONS	TECTONIC SETTING	LOCATION	LOCATION COMMENT
LATITUDE MIN	LATITUDE MAX	LONGITUDE MIN	LONGITUDE MAX
LAND OR SEA	ELEVATION MIN	ELEVATION MAX	DRILL DEPTH MIN
DRILL DEPTH MAX	SAMPLING TECHNIQUE	SAMPLE ID	SAMPLE NAME
ROCK TYPE	ROCK NAME	ROCK TEXTURE	SAMPLE COMMENT
ALTERATION	ALTERATION TYPE	MIN AGE	MAX AGE
GEOL. AGE	ERUPTION DATE	MATERIAL	MINERAL
HOST MINERAL	ANALYSIS TYPE	METHOD	Total:31

Mineral in MetaData

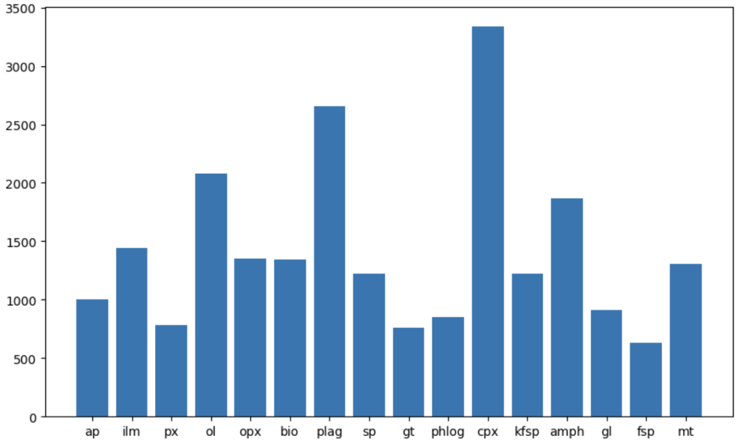
```
In [4]: list(meta_data["MINERAL"].unique())

Out[4]: [nan,
         'OL',
         'CPX; OL; PLAG',
         'OL; SP',
         'CPX; OL; OPX; PLAG; SP',
         'CPX',
         'GL',
         'CPX; OL; PLAG; SP',
         'PLAG',
         'SP',
         'CPX; PLAG',
         'CPX; OPX; PLAG',
         'CPX; OPX',
```

Mineral in MetaData


```
{'LAV', 'CRSP', 'WU', 'PAR', 'FRES', 'BN', 'SILICA', 'LAB', 'CBC', 'JAC', 'EL', 'OLEK', 'F
'NIOB', 'CEB', 'STR', 'HALO', 'TILL', 'COL', 'GEIK', 'SEP', 'MIL', 'CHLO', 'LOW', 'ALT', '
'CLEN', 'DEL', 'SM', 'LOV', 'M', 'PYR', 'LIND', 'REE', 'BIS', 'LAUR', 'HYA', 'FEHYDR', 'ML
', 'RHO', 'DAL', 'RDS', 'CHAM', 'AP', 'MILL', 'OPX', 'ANZ', 'RIP', 'IDD', 'BAO', 'THAU', '
, 'MIA', 'MAFIC', 'SERP', 'FEBO', 'RICH', 'LAU', 'BEI', 'FH', 'MUSC', 'ANM', 'MGHBL', 'TSL
'WAD', 'VISH', 'NAR', 'OK', 'NEP', 'CLH', 'ANT', 'Talc', 'ERI', 'NIG', 'MRN', 'PP', 'NCR',
LM', 'DAC', 'JIN', 'FEMGHBL', 'BY', 'VRN', 'DI', 'MRG', 'CEL', 'THOM', 'RAM', 'MGF', 'AL',
U', 'TON', 'BEN', 'GROSS', 'JD', 'PLG', 'MES', 'HUA', 'STILP', 'FRG', 'TOP', 'PD', 'WIT',
, 'SEL', 'TRID', 'VINO', 'nan', 'PHILL', 'INSOL', 'OL', 'NI', 'FA', 'BRO', 'XEN', 'AK', 'K
', 'MER', 'HLL', 'MGHAS', 'AB', 'ELL', 'LAT', 'CHR', 'CARB', 'SN', 'CRICH', 'SPUR', 'HD',
'PREH', 'BAD', 'LOP', 'HYD', 'CBT', 'RUT', 'SIC', 'SERC', 'UNKNOWN', 'PX', 'CATA', 'KALI',
T', 'HEXSTIB', 'TH', 'ED', 'FE', 'SCHOR', 'MET', 'SHT', 'SP', 'GIT', 'ROSB', 'FEOX', 'BUR'
ANAL', 'PB', 'MAL', 'AMA', 'SAPPH', 'ZKL', 'STRT', 'FSP', 'MATH', 'TUH', 'WAI', 'PYMN', 'M
HR', 'ANK', 'OX', 'APH', 'VS', 'ECK', 'SILIC', 'AMPH', 'SDPH', 'PHLOG', 'GIB', 'QTZ', 'LIC
', 'MONT', 'MGARF', 'TSCH', 'ALAB', 'FEHY', 'MUR', 'LAR', 'RES', 'TA', 'MICA', 'LI', 'FETS
', 'SAL', 'PY', 'PK', 'ANDA', 'L', 'ORC', 'IRT', 'GM', 'MLN', 'KAT', 'PN', 'MOS', 'SEM', '
, 'THOR', 'MCK', 'COV', 'ILM', 'ICB', 'CZ', 'AND', 'SYN', 'SOD', 'GL', 'MGKAT', 'ALUNO', '
'P', 'PARG', 'LAM', 'KUKH', 'GEH', 'HIL', 'GTZ', 'RHÖ', 'VES', 'AN', 'WIN', 'ANTI', 'KASS'
NOS', 'FEHD', 'ROS', 'Opx', 'HIB', 'MNIL', 'AS', 'HUT', 'IR', 'BAZ', 'HYDR', 'MNBAOX', 'UL
'KHI', 'H', 'UMB', 'KHA', 'ASTRO', 'TREM', 'PSLEUC', 'GYR', 'PYRA', 'BRITH', 'PIG', 'MB',
', 'PERR', 'AU', 'FEAU', 'HNG', 'GAH', 'WOD', 'RIN', 'APYR', 'SE', 'NICC', 'PHYL', 'GOL',
UD', 'CYM', 'V', 'PCL', 'FETITS', 'ALSP', 'AUG', 'BER', 'RB', 'GV', 'KAOL', 'WLF', 'ISH',
H', 'NAA', 'RONT', 'MIN', 'POLY', 'TEA', 'MON', 'ALTA', 'KUT', 'TOU', 'SCAP', 'CHLP', 'BA'
MGPIG', 'BRUN', 'COR', 'CLCH', 'PRD', 'CHL', 'CELE', 'SAM', 'DJ', 'MELT', 'MZ', 'TAP', 'GL
'SOL', 'ARAG', 'ALL', 'OSU', 'ZIN', 'FREU', 'QAQ', 'RV', 'F', 'META', 'CRT', 'FO', 'FAS',
AR', 'REIN', 'CORR', 'FSM', 'CHEV', 'JEP', 'PT', 'PARA', 'MAG', 'PEC', 'SYMP', 'MEL', 'HEM
, 'HOLL', 'CUM', 'PARK', 'CRD', 'PER', 'GOE', 'HM', 'HER', 'BETA', 'GED', 'KEL', 'DAV', 'S
ANG', 'ZEO', 'CE', 'TAZH', 'CANC', 'CPX', 'SIL', 'STIB', 'HN', 'YTT', 'BRV', 'SAP', 'AC',
', 'PSB', 'AG', 'MGSD', 'GAL', 'NAT', 'FETS', 'SD', 'CLZ', 'ROD', 'KLP', 'LZ', 'CP', 'MONT
'GIS', 'SID', 'NICK', 'ARM', 'ST', 'DIG', 'SPERR', 'K', 'ZOI', 'UR', 'COF', 'WAR', 'CASS',
AL2SIO5', 'LUE', 'GLAP', 'COMB', 'ZIRC', 'SLF', 'MGS', 'PHOS', 'KIM', 'DIA', 'ANC', 'PYP',
CR', 'CA', 'SHCH', 'MN', 'PLEO', 'AEN', 'GLAUC', 'PUM', 'ABTS', 'PHASE', 'KIMZ', 'VER', 'T
, 'STAU', 'SYL', 'ANORTH', 'STRN', 'BARR', 'GASP', 'GREG', 'MAGH', 'KY', 'PHEN', 'PRS', 'F
LAY', 'SG', 'OLIG', 'RHAB', 'MAGN', 'TIAU', 'OS', 'NOR', 'AEG', 'APO', 'ORE', 'NOON', 'PEF
, 'BAF', 'TALC', 'CC', 'STILB', 'SAPP', 'NE', 'CBL', 'COD', 'AMORPH', 'VLAS', 'GY', 'FLUO'
'BAGH', 'YANG', 'LPM', 'HLR', 'EUX', 'NON', 'ANNITE', 'COSS', 'AGA', 'FLO', 'MICH', 'SI',
, 'AD', 'TIT', 'MULL', 'RNG', 'LEP', 'CORO', 'SMA', 'GRX', 'ARF', 'HBS', 'ZR', 'MAUCH', 'N
, 'HESS', 'KOCH', 'DOL', 'KIRSCH', 'NIO', 'GAI', 'THEN', 'NY', 'ANDE', 'CRIST', 'OXKA', 'Y
```


```
]: 594
```

Mineral in MetaData



Problem formulation

Can we predict if a mineral is relevant for a paper , by looking only into the abstract of the paper?

Problem formulation

can we predict if a mineral is relevant for a paper , by looking only into the abstract of the paper?

- Abstract of the Papers
- MetaData

Implementation Steps

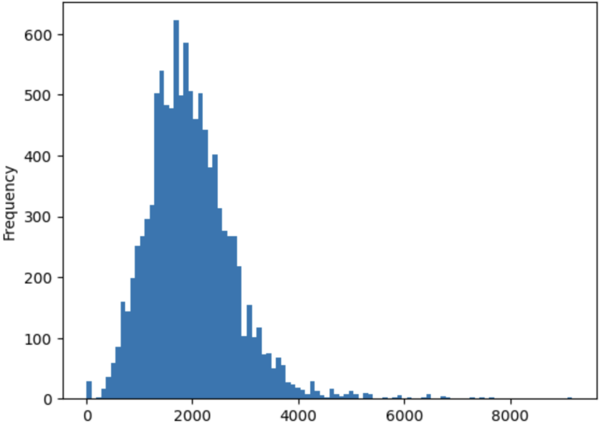
- Data Collection
- Preprocess
- Train
- Evaluation

Collecting Abstracts

```
https://linkinghub.elsevier.com/retrieve/pii/0012821X79901912
https://linkinghub.elsevier.com/retrieve/pii/0012821X78900262
https://linkinghub.elsevier.com/retrieve/pii/0012821X84900736
https://linkinghub.elsevier.com/retrieve/pii/0012821X84900712
https://linkinghub.elsevier.com/retrieve/pii/0012821X83901565
https://linkinghub.elsevier.com/retrieve/pii/0012821X83901553
https://linkinghub.elsevier.com/retrieve/pii/0012821X83901541
https://linkinghub.elsevier.com/retrieve/pii/0012821X8390153X
https://linkinghub.elsevier.com/retrieve/pii/0012821X83901516
https://linkinghub.elsevier.com/retrieve/pii/0012821X83901504
```

- pydoi
- 10656 abstracts for train
- 5000 abstract for eval

Data Analysis



Preprocessing

- Abstracts , and class Data from meta data are concatenated and cleaned.
- Tokenized
- Using glove.840B.300d corresponding embedding for each word is stored .

Word Embedding

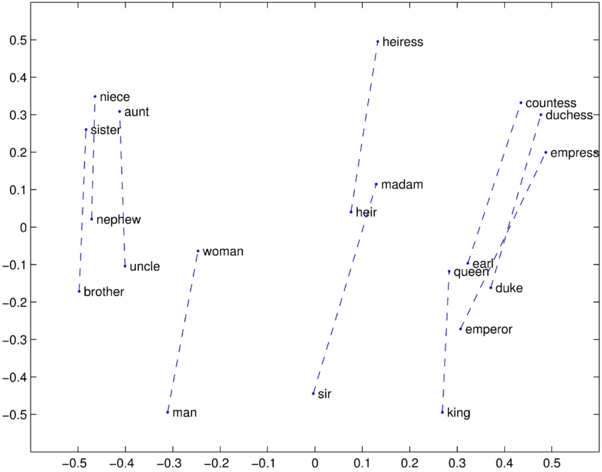
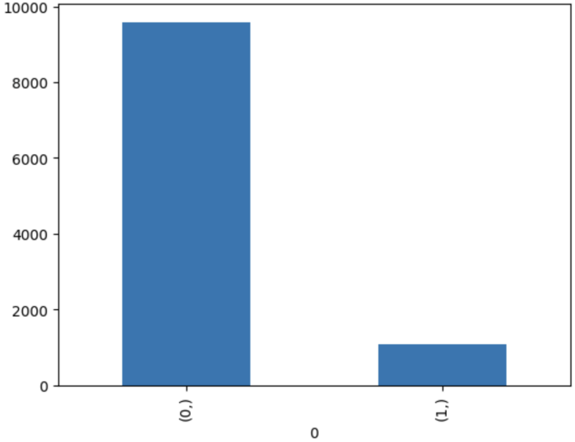


Figure: Given a word , Glove will give you a meaningful vector [PSM14]

Data and class

Feature vector:**word embedding for :abstract+title**
Class:**'cpx'**



model

Text classification using CNN

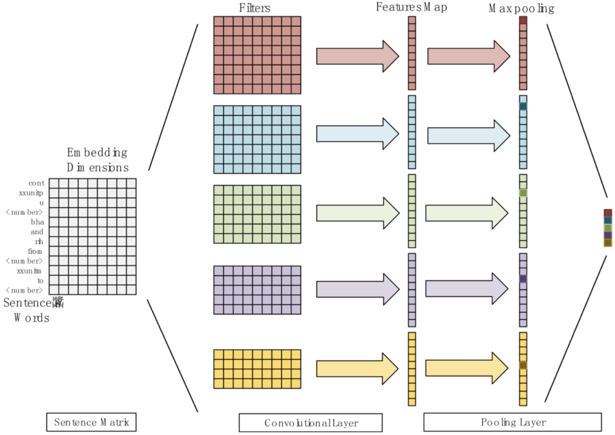


Figure: [HZ18]

Result:Train Accuracy

After Epoch1

ACC	0.954954954954955
tp-rate=(tp)/(tp+fn)	0.7999108734402852
fn-rate=tp/(tp+fn)	0.2000891265597148
tn-rate=tn/(tn+fp)	0.9963147883975273
fp-rate=fp/(tn+fp)	0.003685211602472658

After Epoch 4

ACC	0.9950262762762763
tp-rate=(tp)/(tp+fn)	0.9982174688057041
fn-rate=tp/(tp+fn)	0.0017825311942959
tn-rate=tn/(tn+fp)	0.9941749881122206
fp-rate=fp/(tn+fp)	0.005825011887779363

Result:Evaluation Data

5000 Records

ACC	0.8707627118644068
tp-rate=(tp)/(tp+fn)	0.5735815602836879
fn-rate=tp/(tp+fn)	0.42641843971631205
tn-rate=tn/(tn+fp)	0.9446649029982364
fp-rate=fp/(tn+fp)	0.055335097001763665

Conclusion and immediate steps

- Improving the model with ELmo [Pet+18]
- Including all 594 minerals in a single model
- Predicting other classes in Metadata

Future Work

Multi-Task Identification of Entities, Relations, and Coreferencefor Scientific Knowledge Graph Construction

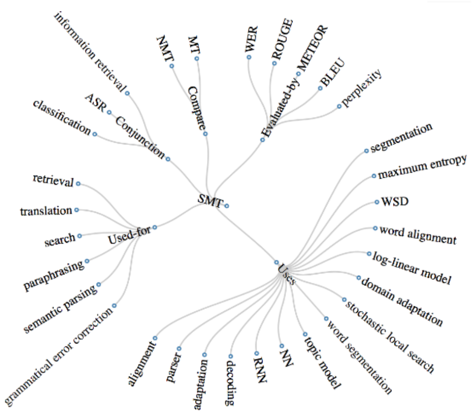
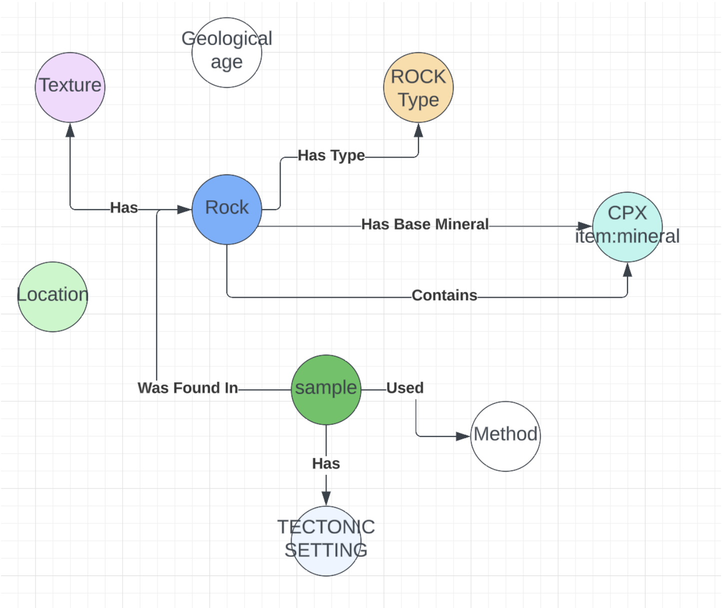


Figure: [Lua+18]

GeoRoc Knowledge Graph



Thanks!

Questions?

Special Thanks to Mathias and Daniel for continued support and feedback.

[PSM14] Jeffrey Pennington, Richard Socher, and Christopher Manning. "GloVe: Global Vectors for Word Representation". In: *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Doha, Qatar: Association for Computational Linguistics, Oct. 2014, pp. 1532–1543. DOI: 10.3115/v1/D14-1162. URL: https://aclanthology.org/D14-1162.

[HZ18] Tzu Fan Hsu and Yaoqi Zhang. "Petroleum Engineering Data Text Classification Using Convolutional Neural Network Based Classifier". In: *International Conference on Machine Learning Technologies*. 2018.

[Lua+18] Yi Luan et al. "Multi-Task Identification of Entities, Relations, and Coreferencefor Scientific Knowledge Graph Construction". In: *Proc. Conf. Empirical Methods Natural Language Process. (EMNLP)*. 2018.

[Pet+18] Matthew E. Peters et al. "Deep contextualized word representations". In: *CoRR* abs/1802.05365 (2018). arXiv: 1802.05365. URL: http://arxiv.org/abs/1802.05365.