Text Categorization

Liu Jin

Problem

- **Classify documents on topics/domains**
- **Dataset: Reuters-21578**
  - 21578 documents
    - 20856 training docs
    - 722 test docs
- **Category sets:**

Category Set	Number of Categories	Number of Categories w/ 1+ Occurrences	Number of Categories w/ 20+ Occurrences
EXCHANGES	39	32	7
ORGS	56	32	9
PEOPLE	267	114	15
PLACES	175	147	60
TOPICS	135	120	57

- **Classification types**
  - Multi-class classification
  - Multi-label classification
- **Case study 1: Classify the 10 most popular topics**
- **Case study 2: Multi-label classification: articles are label by topics and places**

Multi-class classification: Processing and training

- **Data processing**
  - Remove stop words
  - Stemming
  - Tokenization
- **Feature engineering**
  - Count vectorisation
  - TFIDF vectorisation
- **Algorithm**
  - Perceptron
  - Transformer

Results

- **Multi-class classification using Perceptron**

	precision	recall	f1-score	support
acq	0.89	0.95	0.92	77
crude	0.90	0.83	0.86	53
earn	0.71	1.00	0.83	22
ec	0.73	0.94	0.82	17
grain	0.91	0.80	0.85	40
interest	0.89	0.77	0.83	22
money-fx	0.86	0.86	0.86	69
ship	0.85	0.65	0.73	17
trade	0.92	0.84	0.88	58
wheat	0.00	0.00	0.00	0
accuracy			0.86	375
macro avg	0.77	0.76	0.76	375
weighted avg	0.87	0.86	0.86	375

Results

- **Multi-class classification using Transformers**

	precision	recall	f1-score	support
1	0.67	1.00	0.80	22
2	1.00	0.94	0.97	77
3	0.82	0.90	0.86	69
4	0.98	0.85	0.91	53
5	0.86	0.95	0.90	40
6	1.00	0.79	0.88	58
7	0.77	0.77	0.77	22
8	0.89	0.94	0.91	17
10	0.89	0.94	0.91	17
accuracy			0.89	375
macro avg	0.87	0.90	0.88	375
weighted avg	0.91	0.89	0.89	375

Results

- **Multi-label classification using MultiOutputClassifier**
- **Data processing**
  - Each article is labeled by 5 most important topic and 5 most important place (can add other categories, such as Orgs, People)
  - Feature engineering: CountVectorization, TfidfVectorization
- **Algorithm**
  - Sklearn.MultiOutputClassifier
- **Evaluation**
  - Accuracy: 57.5%
  - Hamming loss: 0.06

Sample of dataset

Text	Label
The U.S. Agriculture Department reported the farmer-owned reserve national five-day average price through February 25 as follows (Dlrs/Bu-Sorghum Cwt) - Natl Loan Release Call Avge Rate-X Level Price Price Wheat 2.55 2.40 IV 4.65 -- V 4.65 -- VI 4.45 -- Corn 1.35 1.92 IV 3.15 3.15 V 3.25 -- X - 1986 Rates. Natl Loan Release Call Avge Rate-X Level Price Price Oats 1.24 0.99 V 1.65 -- Barley n.a. 1.56 IV 2.55 2.55 V 2.65 -- Sorghum 2.34 3.25-Y IV 5.36 5.36 V 5.54 -- Reserves I, II and III have matured. Level IV reflects grain entered after Oct 6, 1981 for feedgrain and after July 23, 1981 for wheat. Level V wheat/barley after 5/14/82, corn/sorghum after 7/1/82. Level VI covers wheat entered after January 19, 1984. X-1986 rates. Y-dlrs per CWT (100 lbs). n.a.-not available. Reuter ',	['grain', 'usa']

Outlook

- **Improve accuracy**
  - Parameters tuning
  - Other algorithms
- **Other use cases**
  - Classify by other category sets
  - Classify by hierarchical levels