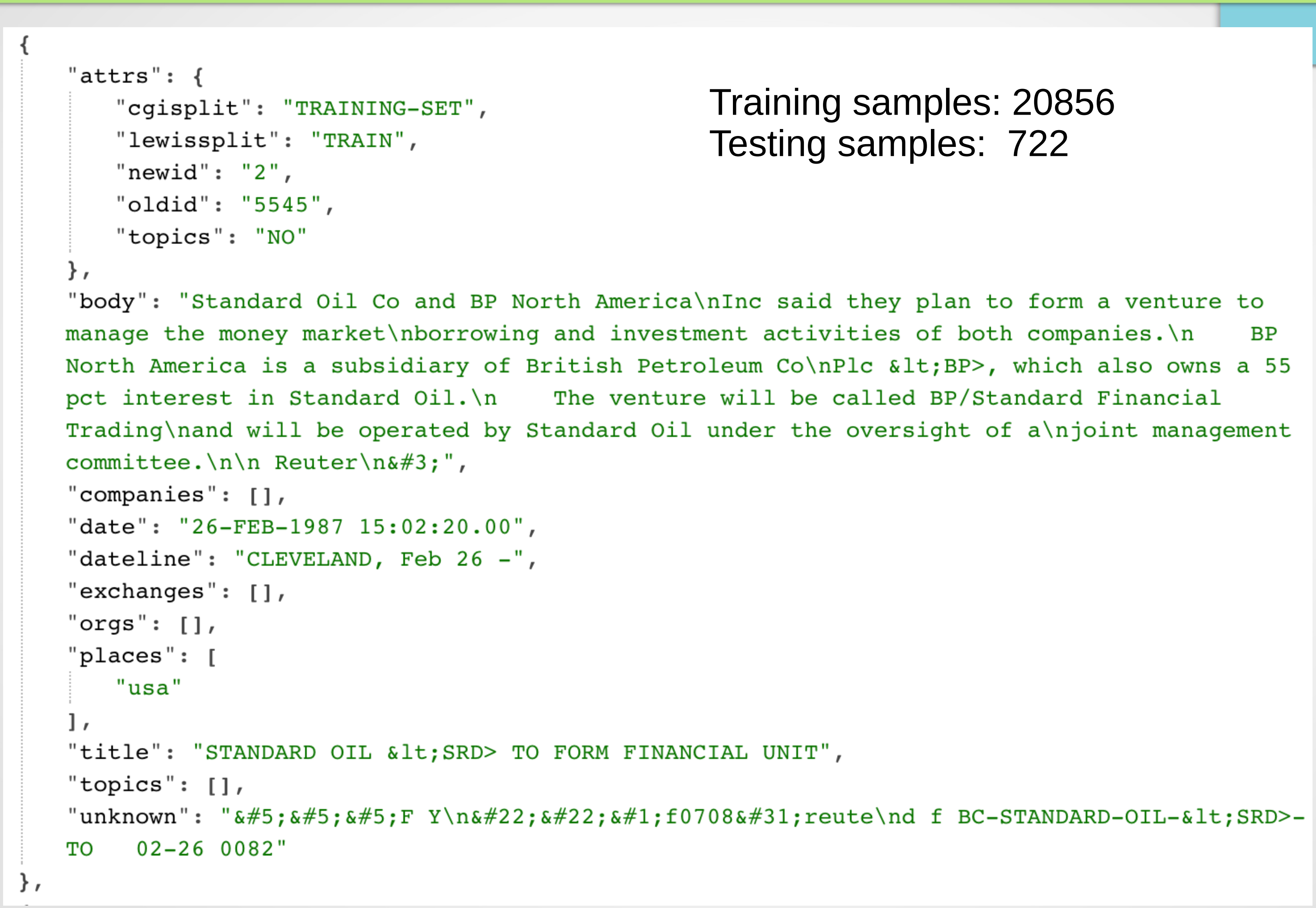
Topics & Places Classifier

Alexandru Mihai,
508

The data. Reuters - 21578

Training samples: 20856
Testing samples: 722



```
{
    "attrs": {
        "cgisplit": "TRAINING-SET",
        "lewissplit": "TRAIN",
        "newid": "2",
        "oldid": "5545",
        "topics": "NO"
    },
    "body": "Standard Oil Co and BP North America\nInc said they plan to form a venture to
    manage the money market\nborrowing and investment activities of both companies.\n    BP
    North America is a subsidiary of British Petroleum Co\nPlc &lt;BP>, which also owns a 55
    pct interest in Standard Oil.\n    The venture will be called BP/Standard Financial
    Trading\nand will be operated by Standard Oil under the oversight of a\njoint management
    committee.\n\n Reuter\n&#3;",
    "companies": [],
    "date": "26-FEB-1987 15:02:20.00",
    "dateline": "CLEVELAND, Feb 26 -",
    "exchanges": [],
    "orgs": [],
    "places": [
        "usa"
    ],
    "title": "STANDARD OIL &lt;SRD> TO FORM FINANCIAL UNIT",
    "topics": [],
    "unknown": "&#5;&#5;&#5;F Y\n&#22;&#22;&#1;f0708&#31;reute\nd f BC-STANDARD-OIL-&lt;SRD>-
    TO   02-26 0082"
},
```

The data

Top 10 topics used in this experiment:

- earn, wheat, money-fx, corn, trade, acq, grain,
- interest, crude, ship

Top 10 places used in this experiment:

- brazil, canada, australia, usa, france, china, uk,
- japan, belgium, west-germany

Topics. Cross validation results

	KNN	Linear SVM (no stopwords)	Linear SVM	Logistic Regression (no stopwords)	Logistic Regression	Logistic Regression (with TFIDF)
earn	0.9480	0.9741	0.9805	0.9751	0.9783	0.9747
wheat	0.9824	0.9905	0.9893	0.9900	0.9902	0.9875
money-fx	0.9645	0.9706	0.9722	0.9737	0.9744	0.9697
corn	0.9821	0.9895	0.9913	0.9884	0.9903	0.9852
trade	0.9686	0.9775	0.9783	0.9746	0.9795	0.9774
acq	0.8917	0.9713	0.9731	0.9732	0.9748	0.9767
grain	0.9669	0.9870	0.9869	0.9865	0.9863	0.9821
interest	0.9688	0.9771	0.9762	0.9745	0.9791	0.9738
crude	0.9687	0.9817	0.9847	0.9635	0.9823	0.9758
ship	0.9716	0.9879	0.9889	0.9860	0.9894	0.9774
SCORE	0.9613	0.9807	0.9821	0.9785	0.9824	0.9780

Places. Cross validation results

	KNN	Linear SVM (no stopwords)	Linear SVM	Logistic Regression (no stopwords)	Logistic Regression	Logistic Regression (+ TFIDF)
brazil	0.9871	0.9912	0.9940	0.9941	0.9941	0.9878
canada	0.9348	0.9778	0.9782	0.9762	0.9786	0.9510
australia	0.9820	0.9914	0.9915	0.9911	0.9913	0.9803
usa	0.8137	0.9048	0.9250	0.8867	0.9301	0.9204
france	0.9777	0.9868	0.9880	0.9879	0.9895	0.9792
china	0.9891	0.9964	0.9962	0.9956	0.9962	0.9926
uk	0.9391	0.9585	0.9670	0.9643	0.9706	0.9520
japan	0.9664	0.9806	0.9836	0.9844	0.9828	0.9746
belgium	0.9873	0.9917	0.9934	0.9920	0.9937	0.9860
west-germany	0.9710	0.9833	0.9865	0.9856	0.9865	0.9765
SCORE	0.9548	0.9762	0.9803	0.9758	0.9813	0.9700

Test results

Topics	Logistic Regression
earn	0.6376
wheat	0.9302
money-fx	0.8292
corn	0.8148
trade	0.7777
acq	0.7009
grain	0.8421
interest	0.5573
crude	0.8118
ship	0.6875
FINAL SCORE	0.7589

Places	Logistic Regression
brazil	1.0
canada	0.8333
australia	0.5833
usa	0.8911
france	0.7555
china	0.9411
uk	0.7333
japan	0.8543
belgium	0.5555
west-germany	0.8732
FINAL SCORE	0.8021