Intelligent Semantic Web Search Service – The Intute Project

Speaker: Yanbo J. Wang

NaCTeM, School of Computer Science
University of Manchester

Project Description

The Intute project, co-funded by JISC (Joint Information Systems Committee) and AHRC (Arts and Humanities Research Council), is a joint work between NaCTeM, Mimas and the Intute Repository Search Project.

The aim of the Intute project is to develop an intelligent semantic web search service using NaCTeM's text mining tools to grant users the benefit of advanced searching within an enhanced subset of the Intute repository, which harvests and aggregates metadata from UK-wide open repositories.

One aspect for the Intute project is to employ the techniques of Text Classification (TC) — automated categorisation of "unseen" documents into pre-defined class-groups.

The Usage of TC in Intute

The "*two-stage*" usage of TC techniques in the Intute project can be detailed as follows.

Stage-one Usage: *Single-label TC*

During the early stages of the Intute project, we are only focusing on those documents belonging to either ***Social Science*** or Bio-medical Science. However, documents in the Intute repository are not necessarily assigned to domain-classes. It is therefore an essential preliminary task to automatically and accurately distinguish these ***Social Science*** or Bio-medical Science documents from other documents in the collection.

Stage-one Usage of TC in Intute

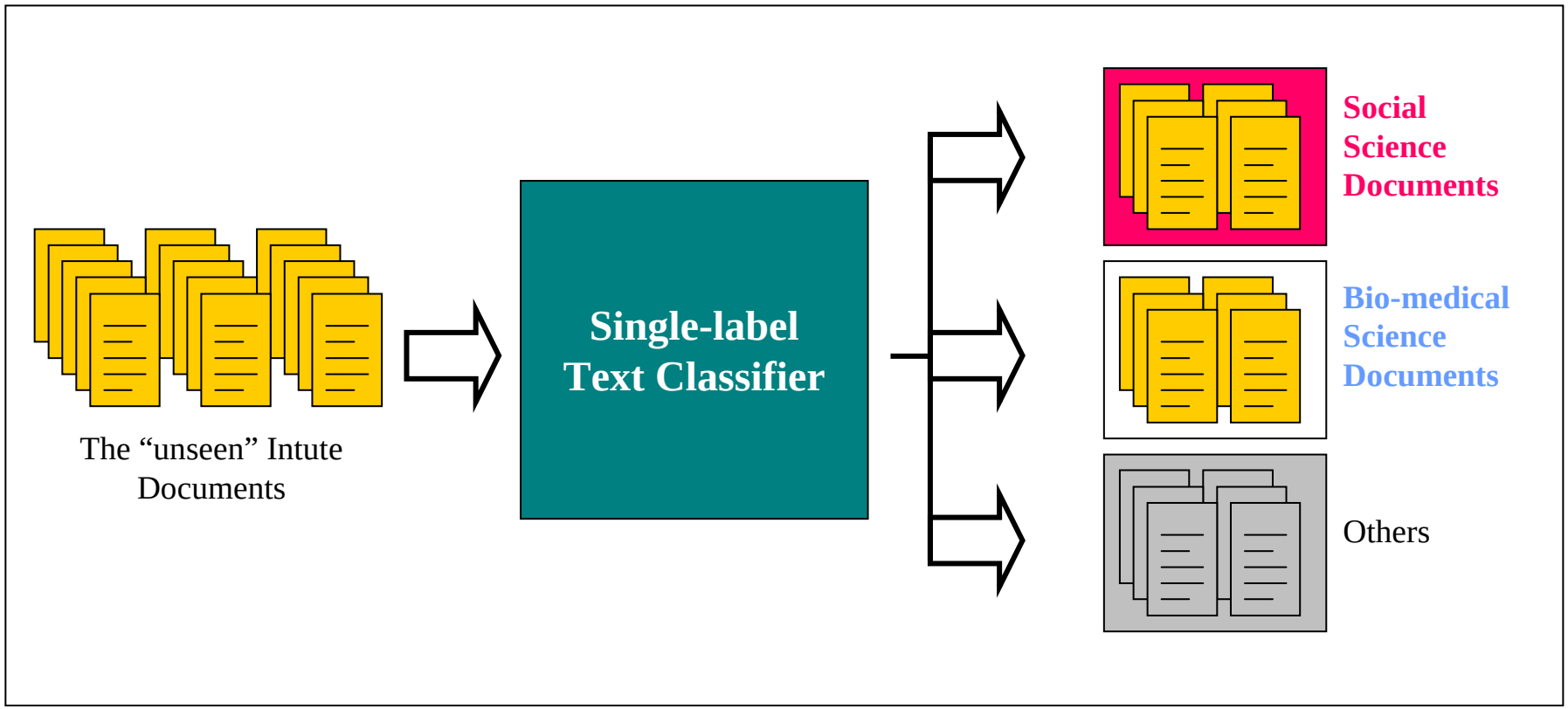
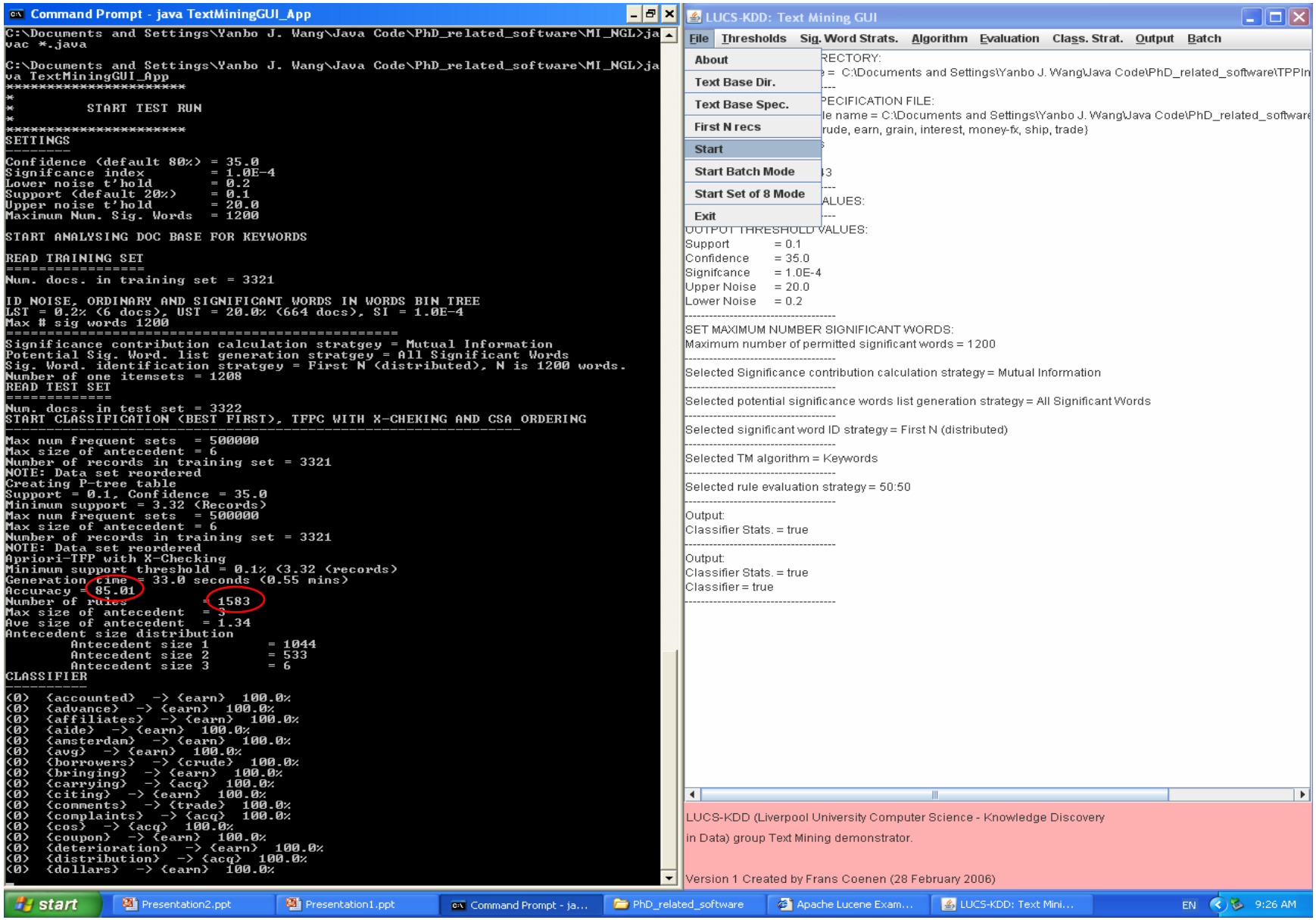


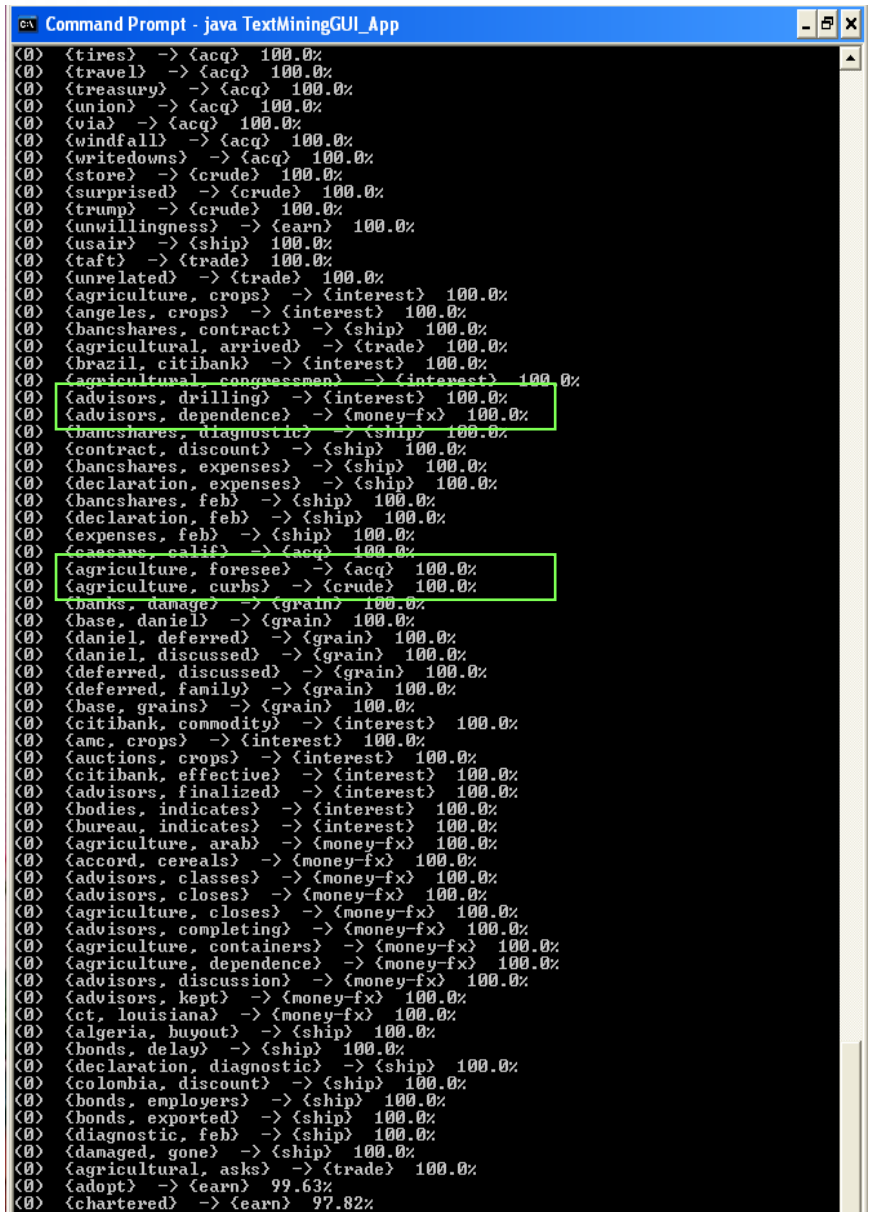
Fig. 1. Stage-one Usage of TC in Intute

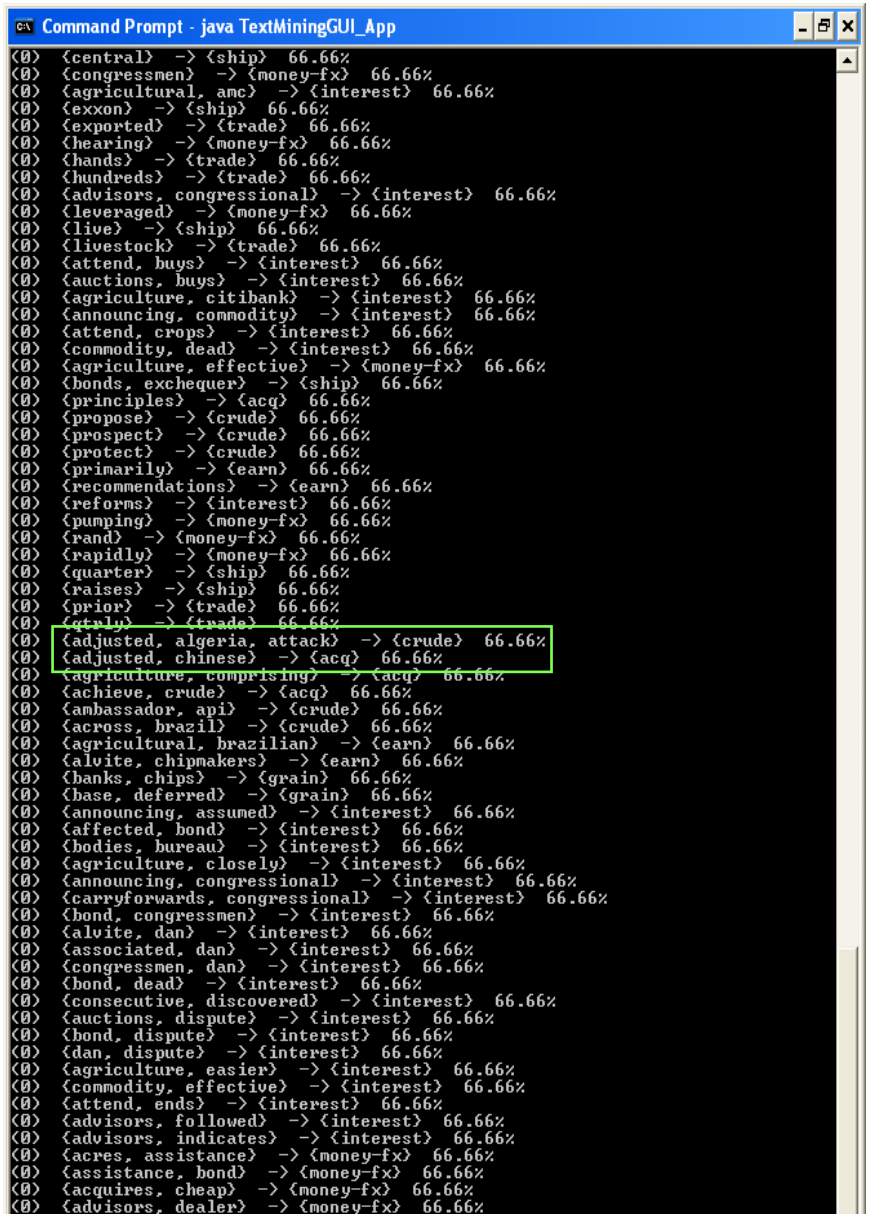
Demo of Single-label TC
– The TFPTC text mining software

Classifier Type	CARM – Classification based on Association Rule Mining
Classifier Name	TFPTC – Total From Partial Text Classification
Document-base	Reuters.D6643.C8
# of Documents	6,643
# of Classes	8, {acq, crude, earn, grain, interest, money-fx, ship, trade}
# of Doc. per Class	{2,108, 444, 2,736, 108, 216, 432, 174, 425}
Feature Selection	Mutual Information
# of Key Words	1,200
Support	0.1%
Confidence	35%
Training : Test	50 : 50

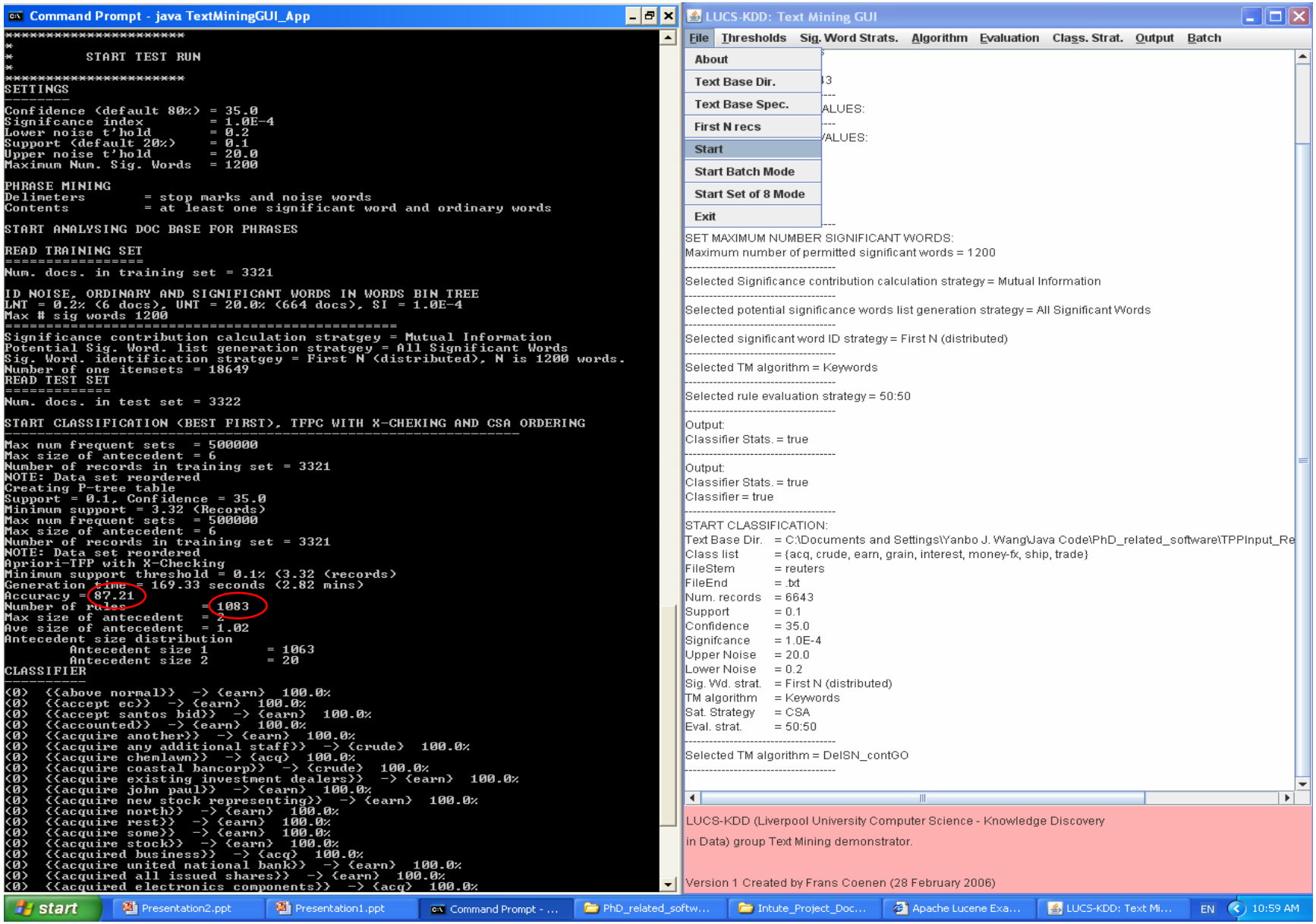


The Keyword-only Approach

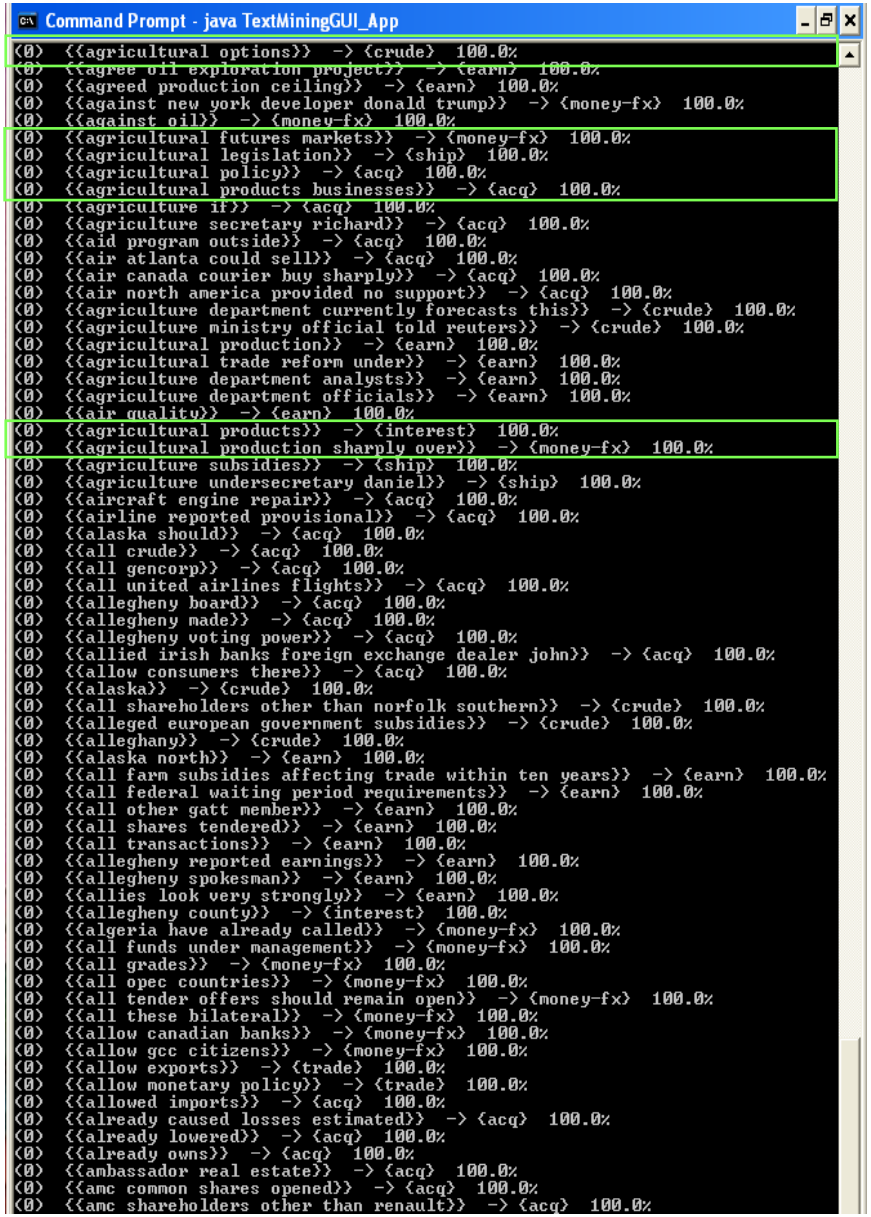




Some Interesting Rules



The Phrase Approach



Some Interesting Rules

Stage-two Usage of TC in Intute

Stage-two Usage: *Multi-label TC*

Usually, a search result is presented as a (long) list of "matching" documents. **Fig. 2** shows the result for querying "*fuel crisis*" on Google. There are total 1,320,000 records returned. Obviously, no one will read them all. Hence presenting this search result in groups, separated by different topics (sub-domain-classes) is suggested.

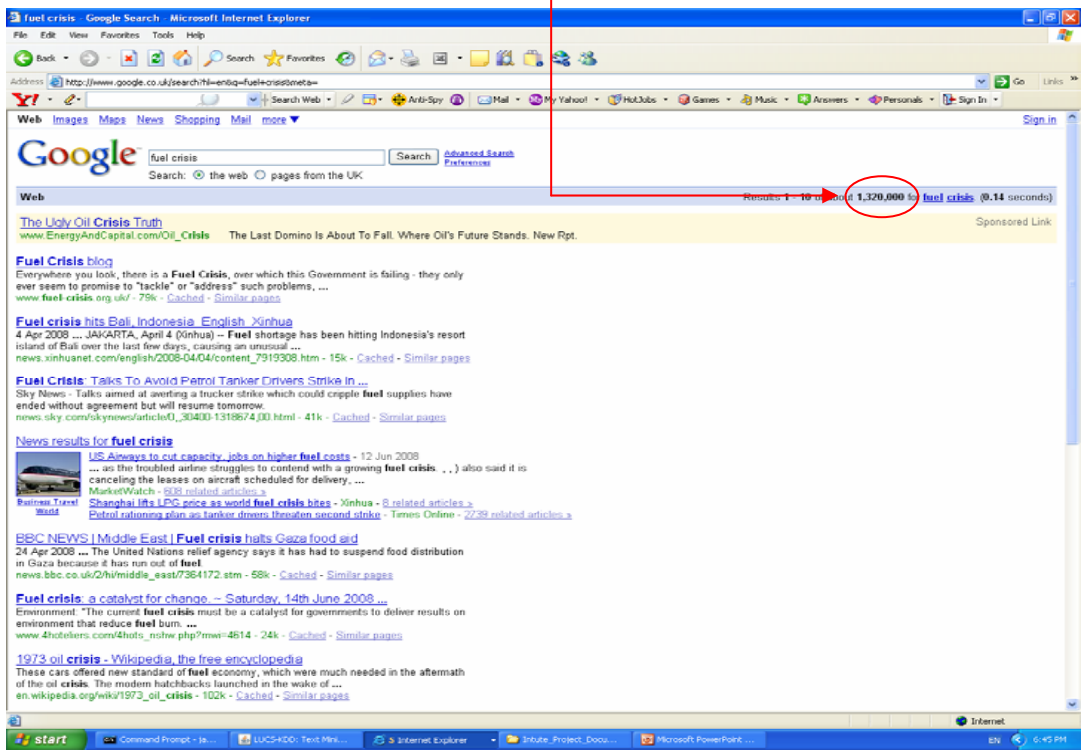


Fig. 2. A Search Result from Google

Stage-two Usage of TC in Intute

Broadly speaking, ***Social Science*** sub-branches include Anthropology, Economics, Education, Geography, History, Law, Linguistics, Political Science, Psychology, Social Work, Sociology, etc. Hence the search result of "*fuel crisis*" can be presented regarding these branch-classes (see **Fig. 3**). Note that a result document (record) may be associated with more than one branch-classes.

Economics	Political Science	Geography	Law
Document # 1	Document # 2	Document # 1	Document # 5
Document # 3	Document # 5	Document # 6	Document # 21
Document # 5	Document # 8	Document # 21	…
Document # 10	Document # 14	…	
…	…		

Fig. 3. Presenting a Search Result in Classes

Strategy of Multi-label TC

From the demo of Single-label TC, we see two rules as follows.

```
(0)  {amc, buys}  -> {interest}  80.0%
(0)  {advisors, comprising}  -> {interest}  80.0%
(0)  {announcing, crops}  -> {interest}  80.0%
(0)  {citibank, crops}  -> {interest}  80.0%
(0)  {advisors, dec}  -> {interest}  80.0%
(0)  {agricultural, decade}  -> {interest}  80.0%
(0)  {bond, decade}  -> {interest}  80.0%
(0)  {dead, effective}  -> {interest}  80.0%
(0)  {bond, fairly}  -> {interest}  80.0%
(0)  {amc, gasoline}  -> {interest}  80.0%
(0)  {arabia, cheap}  -> {money-fx}  80.0%
(0)  {advisors, completes}  -> {money-fx}  80.0%
(0)  {advisors, consistent}  -> {money-fx}  80.0%
(0)  {closes, dependence}  -> {money-fx}  80.0%
(0)  {advisors, fire}  -> {money-fx}  80.0%
(0)  {announcing, follow}  -> {money-fx}  80.0%
(0)  {ct, halt}  -> {money-fx}  80.0%
```

```
(0)  {amc, crops}  -> {interest}  100.0%
(0)  {auctions, crops}  -> {interest}  100.0%
(0)  {citibank, effective}  -> {interest}  100.0%
(0)  {advisors, finalized}  -> {interest}  100.0%
(0)  {bodies, indicates}  -> {interest}  100.0%
(0)  {bureau, indicates}  -> {interest}  100.0%
(0)  {agriculture, arab}  -> {money-fx}  100.0%
(0)  {accord, cereals}  -> {money-fx}  100.0%
(0)  {advisors, classes}  -> {money-fx}  100.0%
(0)  {advisors, closes}  -> {money-fx}  100.0%
(0)  {agriculture, closes}  -> {money-fx}  100.0%
(0)  {advisors, completing}  -> {money-fx}  100.0%
(0)  {agriculture, containers}  -> {money-fx}  100.0%
(0)  {agriculture, dependence}  -> {money-fx}  100.0%
(0)  {advisors, discussion}  -> {money-fx}  100.0%
(0)  {advisors, kept}  -> {money-fx}  100.0%
(0)  {ct, louisiana}  -> {money-fx}  100.0%
```

Hence we indicate that a compound rule can be described as:

$$\{\text{Advisors, Completes/Completing}\} \Rightarrow \{\text{money-fx}\}$$

Strategy of Multi-label TC

Also from the demo of Single-label TC, we see another two rules.

```
(0)  {implement}  -> {trade}  71.42%
(0)  {initially}  -> {money-fx}  71.42%
(0)  {industrialised}  -> {money-fx}  71.42%
(0)  {improving}  -> {earn}  71.42%
(0)  {bilateral, boat}  -> {trade}  71.42%
(0)  {agriculture, fire}  -> {money-fx}  71.42%
(0)  {auditors, discontinued}  -> {money-fx}  71.42%
(0)  {advisors, bonds}  -> {money-fx}  71.42%
(0)  {advisors, fluctuate}  -> {interest}  71.42%
(0)  {announcing, fairly}  -> {interest}  71.42%
(0)  {amc, fairly}  -> {interest}  71.42%
(0)  {announcing, dan}  -> {interest}  71.42%
(0)  {amc, commodity}  -> {interest}  71.42%
(0)  {auditors, bond}  -> {interest}  71.42%
(0)  {agriculture, alvite}  -> {interest}  71.42%
(0)  {arab, deferred}  -> {grain}  71.42%
(0)  {zambia}  -> {trade}  71.42%
```

```
(0)  {fruit}  -> {crude}  68.75%
(0)  {forward}  -> {money-fx}  68.75%
(0)  {foundation}  -> {money-fx}  68.75%
(0)  {gatt}  -> {trade}  68.75%
(0)  {gnp}  -> {trade}  68.75%
(0)  {advisors, citibank}  -> {interest}  68.42%
(0)  {engine}  -> {crude}  68.42%
(0)  {advisors, bond}  -> {interest}  68.0%
(0)  {allegheny}  -> {earn}  67.74%
(0)  {cope}  -> {trade}  67.74%
(0)  {achieve}  -> {earn}  67.62%
(0)  {asks}  -> {earn}  67.56%
(0)  {api}  -> {earn}  67.17%
(0)  {central}  -> {ship}  66.66%
(0)  {congressmen}  -> {money-fx}  66.66%
(0)  {agricultural, amc}  -> {interest}  66.66%
(0)  {exxon}  -> {ship}  66.66%
```

Hence we indicate that a multi-labeled compound rule can be described as:

$$\{\text{Advisors, Bonds/Bond}\} \Rightarrow \{\text{money-fx, interest}\}$$

Further Development

Fig. 4 shows the HASSET (Humanities and Social Science Electronic Thesaurus) categories. The HASSET categories can be used to present ***Social Science*** related documents in subject/domain hierarchies. We introduce an hierarchical multi-label TC problem to map new unlabeled documents to the HASSET hierarchy. This allows the user to concentrate on a "small" group of "interesting" results and offers a solution to the problem of information overload.

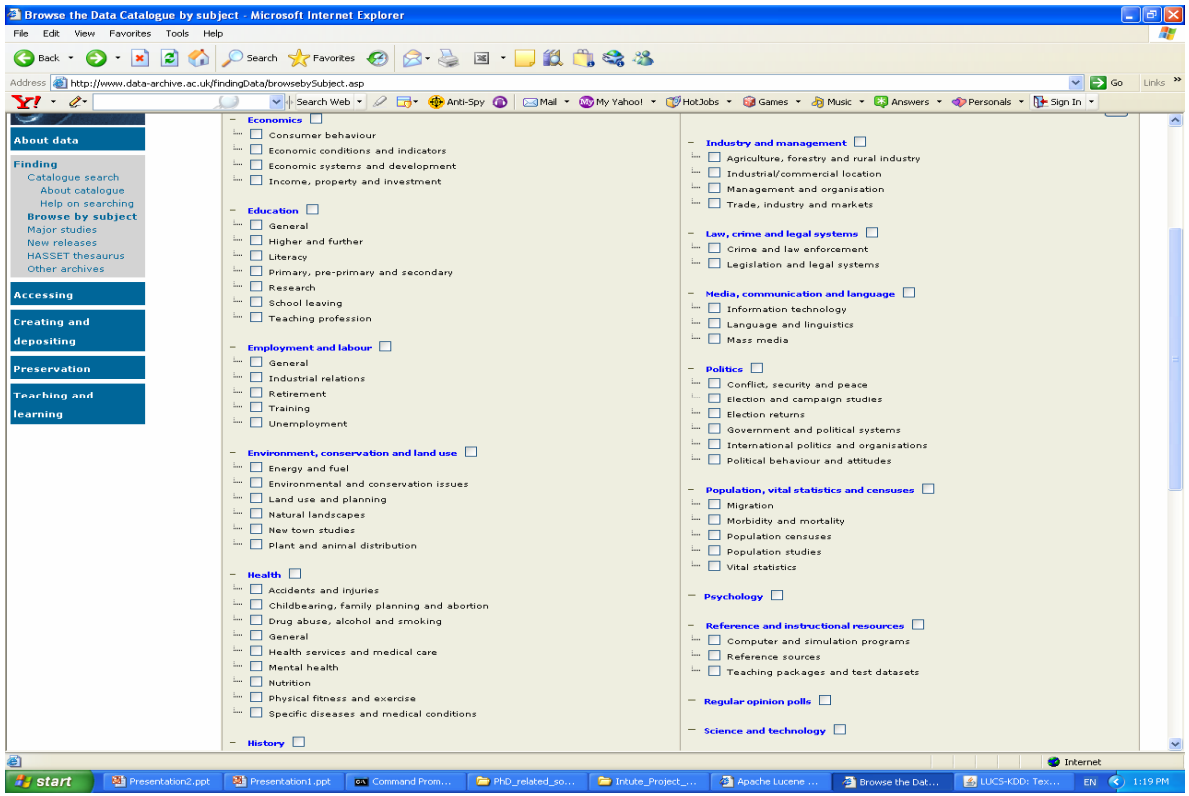


Fig. 4. The HASSET Categories

Summary

The Intute project aims to develop an intelligent semantic web search system that deals with ***Social Science*** and Bio-medical Science documents.

Text classification is a well-known research area that maps documents to pre-defined categories. More than this, the techniques we use allow users to see why those predictions have been made.

As work continues on the Intute project, we will be adding a number of other text mining tools to support cross-repository search focusing on areas of interest to social scientists.

Questions?