Mining Unstructured Data

Outline

1 Course Structure

2 Document Structure and Preprocessing

3 SemEval 2013

Course Structure

The Labs are structured in three main blocks:

- Document Structure and Preprocessing
- DDI and NERC using Machine Learning
- DDI and NERC using Deep Learning

Evaluation

- Lab mark is a 50% of the subject mark.
- All 3 deliverables are weighted the same (33% lab, 16,67% of the subject)
- Each deliverable consists of the code you used for experiments plus a report in pdf format.
- Reports must be maximum 10 pages for each lab task.

Evaluation Criteria

Evaluation Criteria	Score	Excellent	Good	Fair	Poor
Code Effectiveness	**15%**	The code perfectly solves the task, demonstrating a deep understanding of the problem and the NLP techniques used.	The code solves the task with minor errors or inefficiencies. Demonstrates a good understanding of the problem and the NLP techniques used.	The code partially solves the task. Some understanding of the problem and the NLP techniques is demonstrated, but there are significant errors or inefficiencies.	The code does not solve the task. There is little to no understanding of the problem and the NLP techniques demonstrated.
Code Readability and Efficiency	**10%**	The code is extremely well-structured, easy to read, and efficient. It is well-documented with clear, concise comments.	The code is generally well-structured and efficient, with some areas that could be improved. It is mostly well-documented.	The code is somewhat structured and efficient, but there are significant areas for improvement. Documentation is lacking in some areas.	The code is poorly structured and inefficient. There is little to no documentation.
Use of NLP Libraries and Resources	**5%**	The code demonstrates an excellent use of NLP libraries and resources, using them to their full potential to solve the task. The libraries and resources used are appropriate and correctly configured.	The code demonstrates a good use of NLP libraries and resources, but there are some missed opportunities for their use or minor configuration errors.	The code demonstrates a fair use of NLP libraries and resources, but there are significant missed opportunities for their use or significant configuration errors.	The code demonstrates poor use of NLP libraries and resources, with little to no effective use of them or incorrect configurations. The libraries or resources used may not be appropriate for the task.
Analysis and Representation of Results	**45%**	The analysis of the results is thorough and insightful, and the results are represented in a clear, effective manner.	The analysis of the results is generally good, but there are some missed insights. The representation of the results is mostly clear and effective.	The analysis of the results is somewhat shallow, missing significant insights. The representation of the results is somewhat unclear or ineffective.	The analysis of the results is poor, missing most or all key insights. The representation of the results is unclear or ineffective.
Results	**25%**	The score is determined through linear interpolation, where a score of 0 corresponds to the performance of the baseline model, and a score of 10 corresponds to the performance of the model ranked 10th best among the participants.			

Course plan

You can find all relevant information for the course checking the **Course Website**:

- Schedule for each session
- Deadlines
- Source code and data for each task

Course plan for lab sessions

- **Warning 1**: Lab sessions may not be enough. You'll probably need to work on the assignments at home.
- **Warning 2**: Reports must include your conclusions and insights from the experiments. We must know that you understand the experiments you are doing.

Outline

1 Course Structure

2 Document Structure and Preprocessing

3 SemEval 2013

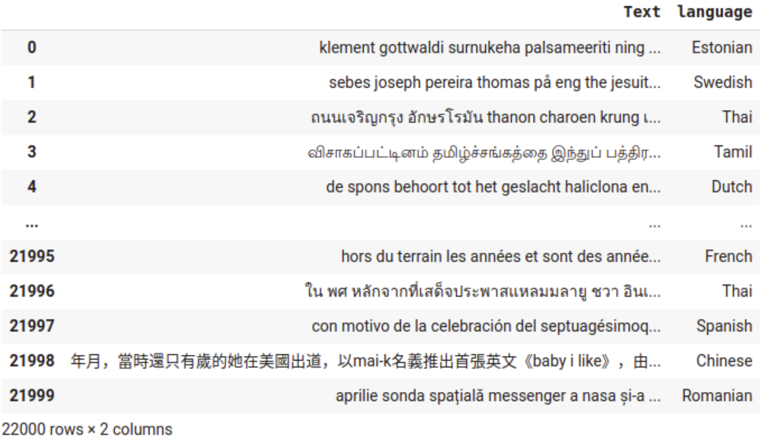
Document Structure and Preprocessing

- Language Detection - Given a text document or sentence, detect in which language it is written.
- 22 languages - Including several different scripts (Arabic, Chinese, Japanese, Latin, etc.)
- Kaggle - Based on the following kaggle notebook:

  https://www.kaggle.com/martinkk5575/language-detection/notebook
- Subtasks
  - Analyze an initial baseline
  - Design your own text preprocessing pipeline and classifier.

Language Detection Dataset

Language detection examples



	Text	language
0	klement gottwaldi surnukeha palsameeriti ning ...	Estonian
1	sebes joseph pereira thomas på eng the jesuit...	Swedish
2	ถนนเจริญกรุง อักษรโรมัน thanon charoen krung เ...	Thai
3	விசாகப்பட்டினம் தமிழ்ச்சங்கத்தை இந்துப் பத்திர...	Tamil
4	de spons behoort tot het geslacht haliclona en...	Dutch
...	...	...
21995	hors du terrain les années et sont des année...	French
21996	ใน พศ หลักจากที่เสด็จประพาสแหลมมลายู ชวา อินเ...	Thai
21997	con motivo de la celebración del septuagésimoq...	Spanish
21998	年月，當時還只有歲的她在美國出道，以mai-k名義推出首張英文《baby i like》，由...	Chinese
21999	aprilie sonda spațială messenger a nasa și-a ...	Romanian

22000 rows × 2 columns

Outline

1 Course Structure

2 Document Structure and Preprocessing

3 SemEval 2013

SemEval-2013 Task 9: DDI Extraction

- SemEval - International Conference on Semantic Evaluation. Several tasks or challenges are posed every edition
- SemEval 2013 Task 9: DDIExtraction - Detect drug names and interactions among them described in text.
- Documents - Documents extracted from DrugBank (Drug description leaflets database) and MedLine (abstracts of medical papers)
- Participants - 11 research teams from around the world (1 Cuba, 1 Italy, 1 Finland, 2 Germany, 1 Portugal, 2 Spain, 3 USA)
- Subtasks
  - Drug name recognition and classification (NERC)
  - Drug-Drug interaction recognition and classification (DDI)

Named Entity Recognition Classifier

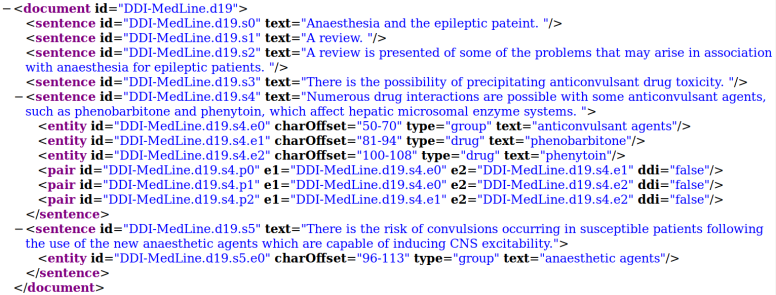
- Given a sentence, classify each of its words into belonging to a named entity (drug) or no.
- Named Entities can be formed by one or several words.
- The dataset includes several types of entities: Brand (Gelocatil), Group (Antipyretic), Drug (Paracetamol), etc.

Drug-Drug Interaction

- Given a sentence with two or more drug names, classify the whole sentence according to the interaction between those drugs.
- Several classes are possible:
  - Effect
  - Advise
  - Mechanism
  - Interaction

DDI Extraction Dataset

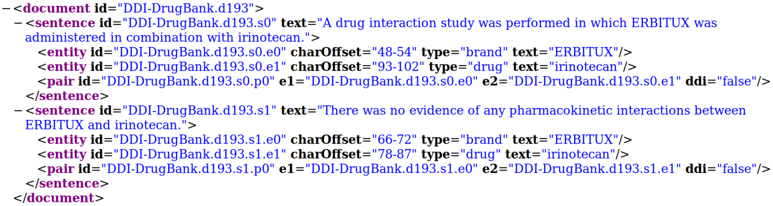
MedLine document example



```
– <document id="DDI-MedLine.d19">
    <sentence id="DDI-MedLine.d19.s0" text="Anaesthesia and the epileptic pateint. "/>
    <sentence id="DDI-MedLine.d19.s1" text="A review. "/>
    <sentence id="DDI-MedLine.d19.s2" text="A review is presented of some of the problems that may arise in association
    with anaesthesia for epileptic patients. "/>
    <sentence id="DDI-MedLine.d19.s3" text="There is the possibility of precipitating anticonvulsant drug toxicity. "/>
  – <sentence id="DDI-MedLine.d19.s4" text="Numerous drug interactions are possible with some anticonvulsant agents,
    such as phenobarbitone and phenytoin, which affect hepatic microsomal enzyme systems. ">
      <entity id="DDI-MedLine.d19.s4.e0" charOffset="50-70" type="group" text="anticonvulsant agents"/>
      <entity id="DDI-MedLine.d19.s4.e1" charOffset="81-94" type="drug" text="phenobarbitone"/>
      <entity id="DDI-MedLine.d19.s4.e2" charOffset="100-108" type="drug" text="phenytoin"/>
      <pair id="DDI-MedLine.d19.s4.p0" e1="DDI-MedLine.d19.s4.e0" e2="DDI-MedLine.d19.s4.e1" ddi="false"/>
      <pair id="DDI-MedLine.d19.s4.p1" e1="DDI-MedLine.d19.s4.e0" e2="DDI-MedLine.d19.s4.e2" ddi="false"/>
      <pair id="DDI-MedLine.d19.s4.p2" e1="DDI-MedLine.d19.s4.e1" e2="DDI-MedLine.d19.s4.e2" ddi="false"/>
    </sentence>
  – <sentence id="DDI-MedLine.d19.s5" text="There is the risk of convulsions occurring in susceptible patients following
    the use of the new anaesthetic agents which are capable of inducing CNS excitability.">
      <entity id="DDI-MedLine.d19.s5.e0" charOffset="96-113" type="group" text="anaesthetic agents"/>
    </sentence>
  </document>
```

DDI Extraction Dataset

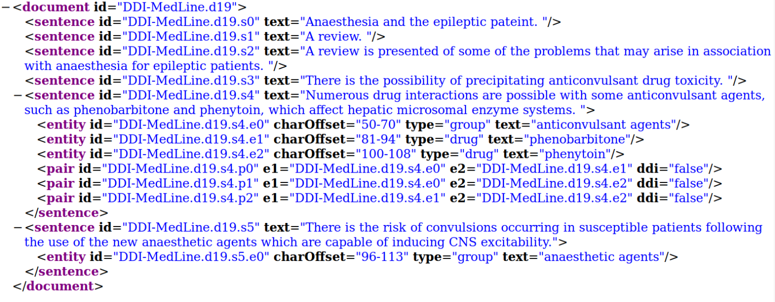
DrugBank document example



```
– <document id="DDI-DrugBank.d193">
  – <sentence id="DDI-DrugBank.d193.s0" text="A drug interaction study was performed in which ERBITUX was
    administered in combination with irinotecan.">
      <entity id="DDI-DrugBank.d193.s0.e0" charOffset="48-54" type="brand" text="ERBITUX"/>
      <entity id="DDI-DrugBank.d193.s0.e1" charOffset="93-102" type="drug" text="irinotecan"/>
      <pair id="DDI-DrugBank.d193.s0.p0" e1="DDI-DrugBank.d193.s0.e0" e2="DDI-DrugBank.d193.s0.e1" ddi="false"/>
    </sentence>
  – <sentence id="DDI-DrugBank.d193.s1" text="There was no evidence of any pharmacokinetic interactions between
    ERBITUX and irinotecan.">
      <entity id="DDI-DrugBank.d193.s1.e0" charOffset="66-72" type="brand" text="ERBITUX"/>
      <entity id="DDI-DrugBank.d193.s1.e1" charOffset="78-87" type="drug" text="irinotecan"/>
      <pair id="DDI-DrugBank.d193.s1.p0" e1="DDI-DrugBank.d193.s1.e0" e2="DDI-DrugBank.d193.s1.e1" ddi="false"/>
    </sentence>
  </document>
```

DDI Extraction Dataset

MedLine document example



```
– <document id="DDI-MedLine.d19">
    <sentence id="DDI-MedLine.d19.s0" text="Anaesthesia and the epileptic pateint. "/>
    <sentence id="DDI-MedLine.d19.s1" text="A review. "/>
    <sentence id="DDI-MedLine.d19.s2" text="A review is presented of some of the problems that may arise in association
    with anaesthesia for epileptic patients. "/>
    <sentence id="DDI-MedLine.d19.s3" text="There is the possibility of precipitating anticonvulsant drug toxicity. "/>
  – <sentence id="DDI-MedLine.d19.s4" text="Numerous drug interactions are possible with some anticonvulsant agents,
    such as phenobarbitone and phenytoin, which affect hepatic microsomal enzyme systems. ">
      <entity id="DDI-MedLine.d19.s4.e0" charOffset="50-70" type="group" text="anticonvulsant agents"/>
      <entity id="DDI-MedLine.d19.s4.e1" charOffset="81-94" type="drug" text="phenobarbitone"/>
      <entity id="DDI-MedLine.d19.s4.e2" charOffset="100-108" type="drug" text="phenytoin"/>
      <pair id="DDI-MedLine.d19.s4.p0" e1="DDI-MedLine.d19.s4.e0" e2="DDI-MedLine.d19.s4.e1" ddi="false"/>
      <pair id="DDI-MedLine.d19.s4.p1" e1="DDI-MedLine.d19.s4.e0" e2="DDI-MedLine.d19.s4.e2" ddi="false"/>
      <pair id="DDI-MedLine.d19.s4.p2" e1="DDI-MedLine.d19.s4.e1" e2="DDI-MedLine.d19.s4.e2" ddi="false"/>
    </sentence>
  – <sentence id="DDI-MedLine.d19.s5" text="There is the risk of convulsions occurring in susceptible patients following
    the use of the new anaesthetic agents which are capable of inducing CNS excitability.">
      <entity id="DDI-MedLine.d19.s5.e0" charOffset="96-113" type="group" text="anaesthetic agents"/>
    </sentence>
  </document>
```