Group project in TDT4215 – Web-intelligence

Spring 2012

		Work form:	Groups Max 3 students
Presentation:	15 minutes presentation		
Hand in:	An electronic copy of both the report and the source code sent to Nafiseh Shabib (shabib@idi.ntnu.no) before 24:00 the 30th of March.	**Delivery:**	1) A written report. Maximal 8.000 words (set the number of words on the title page) 2) Code (electronic copy) 3) Presentation

Assignment

You are a small consultancy providing web search & intelligence applications and solutions to diverse and challenging customers. By chance, you get a call from a small Norwegian (of all places?) publisher of an electronic handbook for pharmaceutical interventions (Legemiddelhåndboka). They want to improve their current keywords-based search interface and provide something that clinicians find efficient and precise. In fact, they want to make their therapy recommendations available in the electronic patient record (EPR) system. However, cluttering the already scarce space in the patient record interface with a web browser and search engine is not attractive, so they want you to design and develop a search system that uses the patient record notes for the current patient and uses it as a search query to the various chapters of the handbook, see Figure 1. The handbook, as well as the patient narratives are written in Norwegian, but you actually consider that a good challenge for making a reasonably language-independent system.

- Download, preprocess, parse and index the handbook from: http://legemiddelhandboka.no/download. Consider how to use the information in the links/anchors, headings and other structure of the handbook. Also take into consideration how to capture the information in the common higher-level chapters.
- Download and preprocess (as necessary) relevant patient cases from those presented as PBL-exercises for students at the medical faculty: https://pbl.medisin.ntnu.no/.

Two classification systems are of particular relevance as a source of terms and also for terminological (ontological…) reasoning.

- International Classification of Diseases, version 10 (ICD-10), described here: http://www.who.int/classifications/icd/en/and available in Norwegian as an OWL2 file here: http://bit.ly/ws8ebb. Note that the Norwegian version

contains links to counterindications, not originally part of the international ICD-10.

- Anatomical Therapeutic Chemical classification of Drug (substances) is described here: http://www.whocc.no/atc/structure_and_principles/. A Norwegian version is available from the its-learning repository, as both a Prolog fact file and as an XML-file.

Tasks

1. Autocoding ICD-10

a. For each sentence in the clinical notes, find relevant ICD-10 codes based on match between the description part of the ICD-10 codes (see the ICD-10 ontology) and the sentences in the clinical notes. A sentence can match zero to many ICD-10 codes, but preferable no more than one – one that is as specific as possible. To store the relation between a sentence and ICD-10 codes, make a table/list where each row contains the document number, sentence number, followed by the relevant ICD-10 codes (ranked).

Clinical note	Sentence	ICD-10
1	1 2 3	M4.2, P45.1, P45.2 M3 T68.7
2	1 2 3	. . .

Example: Given the sentence:
12, 3 Pasienten har smerter i korsryggen, trolig Hekseskudd.
By searching in the ICD-10 ontology the system finds that *M54.5* is a good match. As a result M54.5 should be added as a relevant ICD-10 code for sentence 12, 3 in the results list.

b. Do the same as above for the therapy chapters in Legemiddelhåndboken (it is roughly organized in the same hierarchical fashion as ICD-10). Note that the chapters usually contain sub-chapters, and some even sub-sub-chapters. You will only need to code the (sub(sub)) chapters that contain text.

Chapter	Sentence	ICD-10
T12.1.1	1 2 3	R1.1 M2, A8.1 T68.7, T68.9

T12.1.2	1 2 3	. . .

2. **Autocoding ATC**
   a. and b. Do the same as in Task 1, but with the ATC-classification instead of ICD-10. Store the results in the same way you stored the results from Task 1.

Clinical note	Sentence	ATC
1	1 2 3	QG51AA01 . .
2	1 2 3	. . .

3. **Ranking using vector models**

   In this task you are going to use an information retrieval (IR) method for extracting and comparing (latent) semantic information in documents, based on statistical information. We suggest that you use a vector *space* model (VSM) based method called Latent Semantic Indexing (LSI).

   - The task is to calculate similarities between clinical notes and therapy chapters in Legemiddelhåndboken. For each clinical note, belonging to a certain patient case, make a ranked list of therapy chapters in Legemiddelhåndboken, similar to Figure 1. To achieve this you will need to create a term-by-document matrix, where the documents are both the clinical notes and the therapy chapters in Legemiddelhåndboken.

4. **Evaluation**

   Evaluate your results from Task 3. Propose technique/methods for evaluating the results automatically. Write a short report (about **?** pages).

5. **Exchange evaluations**
   - Exchange the reports you wrote in Task 4. Every group sends their report by email to the responsible und.ass, who passes them on to the other groups.
   - For each unique technique/method found in the other reports, write a short evaluation.

6. **Improving the ranking**

   Here you are to improve the results from Task 3, first by using the results from Task 1 and 2. You should now have the relevant codes for each sentence in both the clinical notes and Legemiddelhåndboken.

a. For each clinical note (or patient), use only the codes from Task 1 and 2 to select and rank relevant chapters in Legemiddelhåndboken. Here you should exploit the hierarchical structure of the ontologies to calculate similarities.

b. Combine the results from Task 6.a with the results from Task 3. You will have to find a way to weight the different rankings given by the two approaches.

7. Gold Standard

We will provide a "gold standard" of patient cases (clinical notes matched with relevant codes). Evaluate your different ranking methods with this gold standard.

8. Possible improvements

Propose ways to extend and improve the methods.

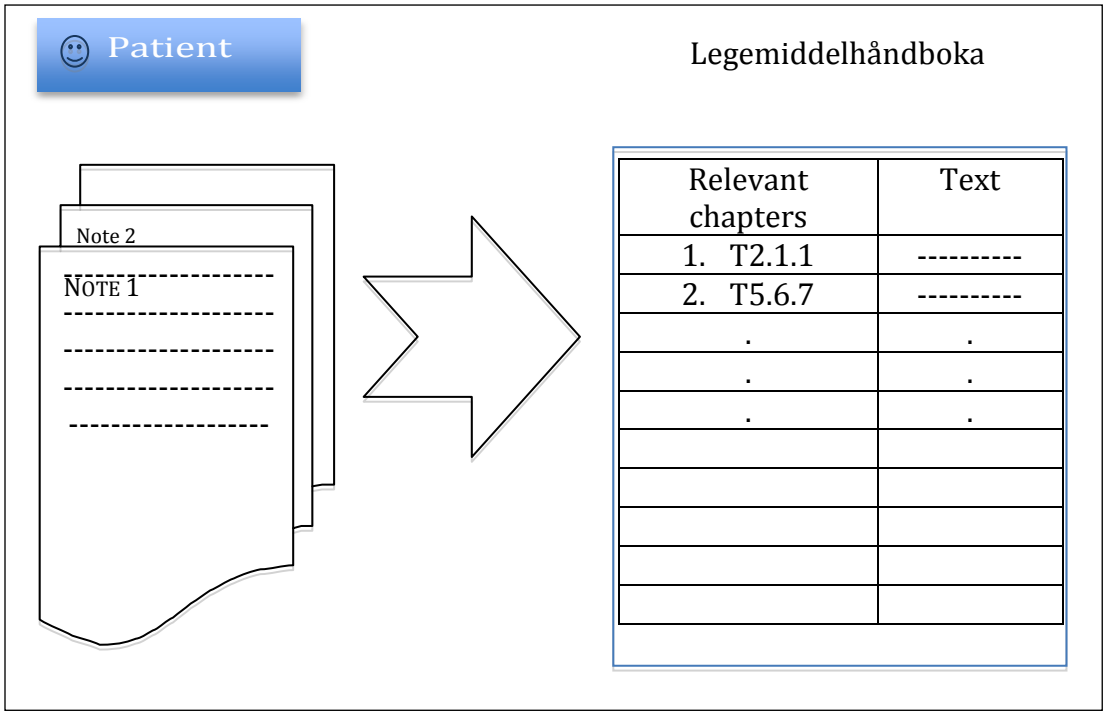

Figure 1.

Resources

- You may use tools like GATE or NLTK, and for example a search engine like Lucene.
- Only approved libraries can be used. These libraries will be found on the home page of this course. If you would like to use some other libraries, then

these libraries will have to be approved first. If approved, they will be made available on the homepage of this course for others to use as well.

Report

- Introduction.
- System architecture.
- Components; function and role. Show the structure of the components (use e.g. UML).
- Limitations given in the assignment.
- Present theories used. Are any theories outside of the curriculum been used? Explain the reason for the selection.
- How have the theories been adapted to the assignment?
- Summary of the results.
- Evaluation and discussion of the results.
- Thoughts of potential improvements.
- Conclusions.

Presentation

- Explain the system architecture and function and role of the components.
- Which components did you make yourself?
- Present and discuss the results of algorithm.
- Discuss (pros and cons) Explain the selected classification algorithm.
- Evaluation of the classifiers.

Note that you can find a stored data file in

its learning -> Data files -> act -> "decomposed codes with names" with a set of Prolog facts on the form:

```
atcname( [a,1], 'munn- og tannmidler', 1, 1).
atcname( [a,1,a,a], 'midler mot karies', 1, 1).
atcname( [a,1,a,a,1], 'duraphat', 4, 1).
atcname( [a,1,a,a,1], 'fluorette', 4, 1).
atcname( [a,1,a,a,1], 'flux', 4, 1).
atcname( [a,1,a,a,1], 'natriumfluorid', 1, 1).
atcname( [a,1,a,a,1], 'xerodent', 4, 1).
atcname( [a,1,a,a,2], 'natriummonofluorfosfat', 1, 0).
atcname( [a,1,a,a,3], 'olaflur', 1, 0).
atcname( [a,1,a,a,4], 'tinnfluorid', 1, 0).
atcname( [a,1,a,a,30], 'kombinasjoner', 1, 0).
atcname( [a,1,a,a,51], 'natriumfluorid, kombinasjoner', 1, 0).
atcname( [a,1,a,b], 'antiinfektiva og antiseptika til lokal behandling i munn', 1, 1).
```

The act-code is represented as a list of sub-codes for the different levels, the term is either substance name or product name, and the last two places can be ignored. Should be simple to parse, - and even simpler to use as Prolog facts if needed :-)

You should start with the following cases from the MFEL1010 course:
4. Trond Øvrebotten (HJERTEINFARKT)
5. Hanne Kristoffersen (INFEKSJONER)
6. Pasient 56 år (OBSTRUKTIV LUNGESYKDOM)
7. Geir Eriksen - skoleelev (DIABETES)

You should cut out the relevant Norwegian text. We´re pointing you to the entire database just to let you see the full context of the cases, and the corresponding english versions"