Possible Exam Questions LSKD

Introduction to Life Science Knowledge Discovery

1. what is the definition of knowledge discovery according to Fayyad?
2. sketch the knowledge discovery process as a cartoon
3. what is a “model”? give a short definition of a “model” and if possible provide some examples of different classes of models
4. give a short definition of the following terms: synonym, homonym, ambiguity, cooccurrence
5. draw a confusion matrix and explain the following terms: recall, ,precision, sensitivity, specificity, coverage
6. what does “URI” mean? Give a simple example of an URI!
7. What is the concept behind “linked open data” (LOD) ?
8. What are knowledge representation formats that you are aware of?
9. What is the problem with the syntactic web? What is the difference between syntactic and semantic web?
10. What is a possibility to add semantic information to web content? (-> Ontologies!)
11. What are three different languages for “explicit specifications” of ontologies? (Graphical notations (1), Logic based (2), probabilistic/fuzzy (3))
12. What does “RDF” stands for? What is the concept behind RDF and what can you do with this structure?
13. What is the RDF data model? Give a very simple example!
14. What are the main problems dealing with RDF?
15. What does “OWL” stands for? What are the main properties of OWL?
16. How can you store these kinds of data efficiently? (Triple Store)
17. What are the main advantages of representing data in Triple Stores?
18. Examples of biological data stored as Triple Stores?

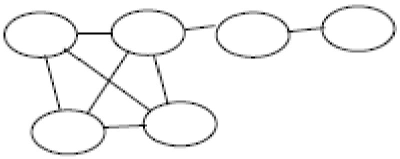
Biological knowledge and its representation

19. Explain the basic construction principles of MeSH. DO NOT write what the acronym MeSH stands for. Just write down what the fundamental building principles are. Please emphasize on the Descriptor/Concept/Term structure.
20. How would you build an ontology? Sketch your approach to construct a knowledge representation describing Alzheimer´s disease
21. Describe the content of UMLS and explain, why UMLS is called a “meta-thesaurus”

Fundamental methodology for knowledge discovery

22. What are the three fundamentally different machine learning approaches?
23. Which of these three approaches do you apply when only unlabelled data is available?
24. Describe the Vector Space model
25. Describe Naive Bayes, what are the assumptions?
26. Explain the following concepts: Instance, Attribute, Feature, Class
27. What is the difference between a nominal, ordinal, or numerical attribute/feature? Explain!
28. What is exploratory data analysis? Mention three artefacts you could observe and how you would deal with them.

29. What are the four steps in the data mining workflow, explain in short the basic idea of each step.
30. From exploratory data analysis you can learn which data preprocessing steps could be beneficial for training a model. Which data transformation is essential when using nearest neighbour for learning and inference?
31. What is the difference between lazy and eager learners? Give one example for a lazy and one example for an eager learner!
32. How can nearest neighbour be used for regression?
33. Give the algorithm of building a decision tree as pseudo code!
34. What is the idea of information gain? Explain in comparison to the term entropy!
35. What is the fundamental idea of Naïve Bayes?
36. Given p(y, x1, x2, x3) as probability distribution, where y is the class to be predicted, and xi, i: {1,2,3}, are the features. What is the factorization typically applied in Naïve Bayes?
37. The Naïve Bayes can be understood as a directed probabilistic model. Draw the model from the Question above.
38. You trained a model and you think it could be overfitted:
    a. How can you prove that?
    b. What can you do to make the model more generalizable?
39. Give the algorithm of k means clustering as pseudo code!
40. How can you measure the quality of clusters?
41. You seed in k means seems to be no good? What could you do?
42. What is the difference between precision and accuracy?
43. You have a classification problem with 100 instances of class A and 1000 instances of class B:
    a. Your accuracy is 90%. Are you happy? Explain!
    b. The F1 measure of class B is 90%. Are you happy? Explain!
    c. The F1 measure of class A is 90%. Are you happy? Explain!
44. What is the GiGo principle? What do you do if you assume that it is the reason for a not-so-good classifier?
45. You get 10000 instances for a task with two classes and 5 features. What are your next steps? Discuss possible problems!
46. You get 1000 instances for a task with 2 classes and 100000 features. What are your next steps? Discuss possible problems!
47. How could you check empirically if your data is probably sufficient?
48. What is the Bias Variance Dilemma? Explain!
49. Which two different approaches do you know for feature selection? What are advantages and disadvantages?
50. What is Occam's Razor?
51. What is the vector space model?
52. What is the idea of tf.idf? What for do you use it?
53. Given the Markov Graph, draw a possible factor graph and explain your choice!



54. What are advantages of factor graphs instead of the representation of undirected markov graphs (or dependency graphs)?
55. What is named entity recognition?
56. Which different approaches do you know for Named Entity Recognition?
57. Which different approaches do you know for relation extraction?
58. Which features are commonly applied in relation extraction? Give examples and explain!
59. What is the difference between the factorization of a probability distribution when the goal is an undirected graphical model in comparison to yielding a directed graphical model?

60. What is the simplest way to specify a probability distribution?

61. Explain the following concepts:

a. **Independent Variable**: “The independent variable is typically the variable representing the value being manipulated or changed and the dependent variable is the observed result of the independent variable being manipulated.” (there was nothing in the slides, so I asked Wiki: http://en.wikipedia.org/wiki/Independent_variable)

b. **Regression**: supervised learning of a numerical predictor, also: curve fitting (lecture 03, slide 6)

c. **Discretisation Artefact**: “In mathematics, discretization concerns the process of transferring continuous models and equations into discrete counterparts.” (again no explanation in the slides, definition from Wiki: http://en.wikipedia.org/wiki/Discretization)
related to that term: **Jittering** – “Jittering is the act of adding random noise to data in order to prevent overplotting in statistical graphs. Overplotting can occur when a continuous measurement is rounded to some convenient unit. This has the effect of changing a continuous variable into a discrete ordinal variable.” (lecture 03, slide 22; http://blogs.sas.com/content/iml/2011/07/05/jittering-to-prevent-overplotting-in-statistical-graphics/)

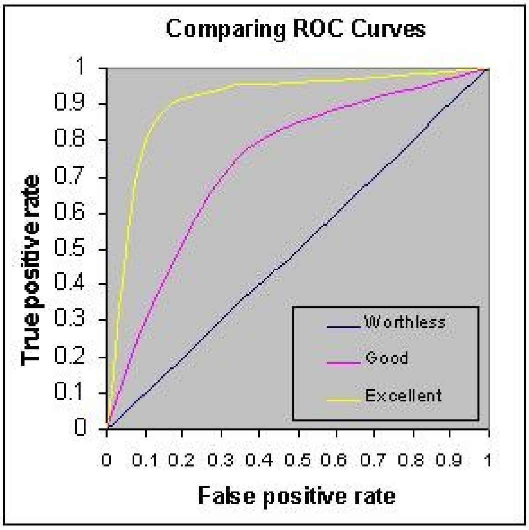
d. **GiGo Principle**: “Garbage in - garbage out”, self-explaining I guess. Also known as the slogan of b-it. (lecture 08, slide 26)

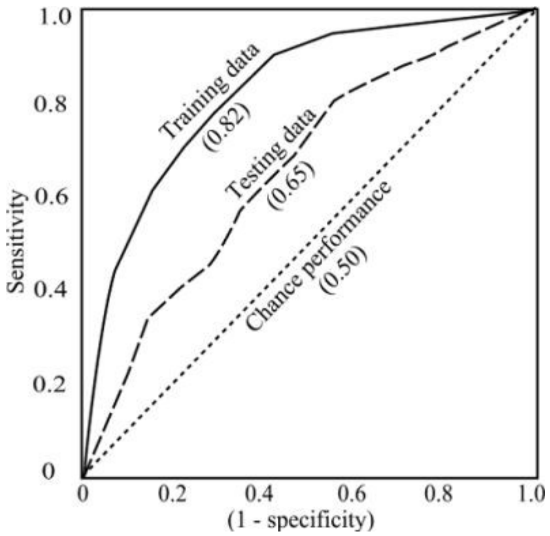
e. **Occams Razor**: “make it as simple as possible, but not simpler” (lecture 08, slide 26)

62. Describe different types of Data mining questions: (lecture 03, slide 5)
- **supervised**: data given with annotation; classification/categorization, regression/curve fitting, structured learning (lecture 03, slide 6)
- **unsupervised**: find patterns in data; clustering, semi-supervised learning (lecture 03, slide 8)
- **reinforcement**: data with positive or negative feedback; learn to play a game, balance a stick, robotic movements (lecture 03, slide 7)

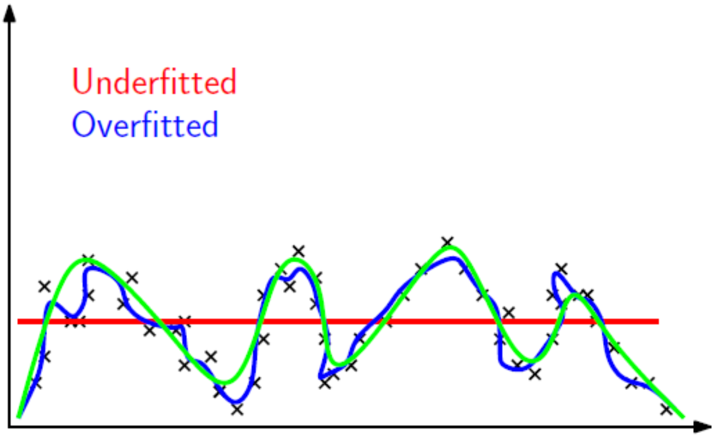
63. What is Information Retrieval?
select those sources (e.g. documents) which contain relevant information (lecture 05, slide 5)

64. What is the ROC-plot, what can we see from it?
I think we didn’t cover it; there is nothing in the slides. Following definition from Wiki:
“In signal detection theory, a receiver operating characteristic (ROC), or simply ROC curve, is a graphical plot which illustrates the performance of a binary classifier system as its discrimination threshold is varied. It is created by plotting the fraction of true positives out of the positives (TPR = true positive rate) vs. the fraction of false positives out of the negatives (FPR = false positive rate), at various threshold settings. TPR is also known as sensitivity, and FPR is one minus the specificity or true negative rate.”
(http://en.wikipedia.org/wiki/Receiver_operating_characteristic)

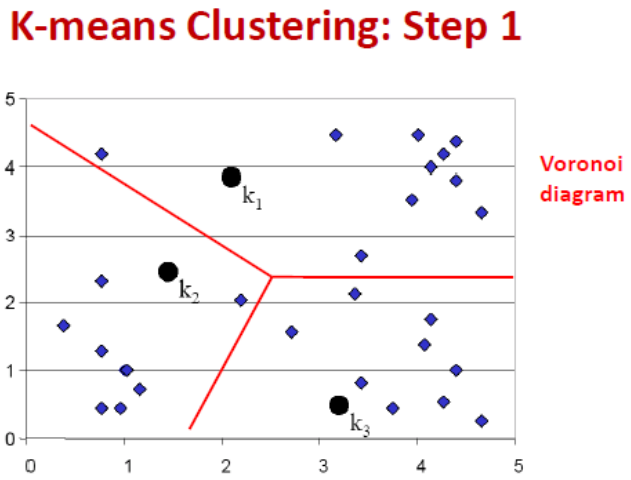




65. What is overfitting/underfitting? (lecture 08, slide 27)
    Overfitting: being too specific/complex; fitting the data too good
    Underfitting: being too general; not fitting the data at all



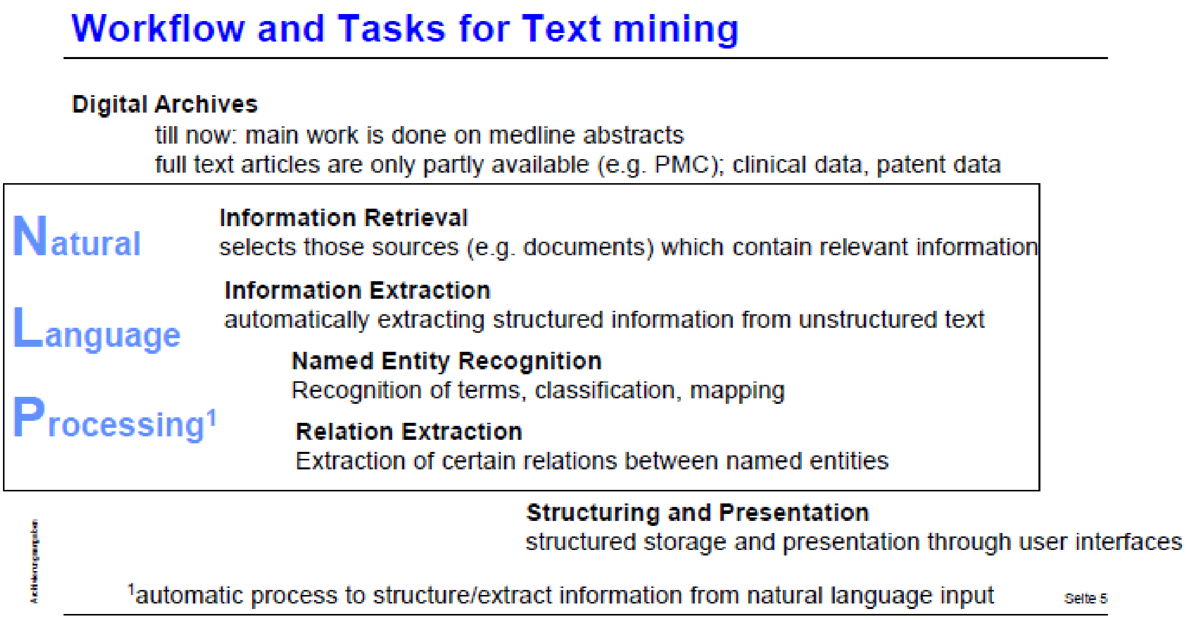
66. Describe the k-Means clustering algorithm. What is the potential usage?
    Predefinition of k clusters; initialization of cluster centers for each cluster; assignment of each instance to its nearest cluster center; re-computing of the cluster centers; iteration until convergence (lecture 08, slide 13)
    potential usage: optical character recognition (OCR), speech recognition, image analysis (various internet sources)


K-means Clustering: Step 1

67. What is Exploratory Data Analysis, given a set of patient data with the attributes (age, gender, disease state, APOE allele variant, systolic blood pressure) describe how you will proceed.
    problem definition and data acquisition; conversion of data into right format; do exploratory data analysis; check descriptive statistics (mean, median, quartiles…) and plots (histograms, correlations, scatter plots, boxplots…)
    caution with: discretization artifacts (see question 61c), outliers, distributions/ ranges/ correlations, balancedness of data, missing/constant values, necessary transformations (lecture 03, slide 21)
    For the given task I would plot the distributions and ranges of each variable and the correlations between some of them (e.g. age or gender with systolic blood pressure, disease state with APOE allele variant etc.). This will help identifying outliers and finding relationships between variables.

Text mining technology

68. Explain the following text mining terms:



Workflow and Tasks for Text mining

Digital Archives
till now: main work is done on medline abstracts
full text articles are only partly available (e.g. PMC); clinical data, patent data

Natural **L**anguage **P**rocessing[1]

Information Retrieval
selects those sources (e.g. documents) which contain relevant information

Information Extraction
automatically extracting structured information from unstructured text

Named Entity Recognition
Recognition of terms, classification, mapping

Relation Extraction
Extraction of certain relations between named entities

Structuring and Presentation
structured storage and presentation through user interfaces

[1] automatic process to structure/extract information from natural language input

Seite 5

NLP: Natural Language Processing; automatic process to structure/extract information from natural language input (lecture 05, slide 5); field between computer science and linguistics, deals with handling language with automated methods, overlaps with text mining (lecture 01, slide 27); conversion of text or spoken language into digital format (lecture notes); “interactions between computers and human (natural) languages”, “Specifically, it is the process of a computer extracting meaningful information from natural language input and/or producing natural language output.” (Wiki: http://en.wikipedia.org/wiki/Natural_language_processing)
Information extraction: automatically extracting structured information from unstructured text (lecture 05, slide 5), “Information extraction (IE) is the task of automatically extracting structured information from unstructured and/or semi-structured machine-readable documents. In most of the cases this activity concerns processing human language texts by means of natural language processing (NLP).” (Wiki: http://en.wikipedia.org/wiki/Information_extraction)
Named entity recognition: Recognition of terms, classification, mapping (lecture 05, slide 5), classifying terminology, term annotation, normalisation (lecture notes); “Named-entity recognition (NER) (also known as entity identification and entity extraction) is a subtask of information extraction that seeks to locate and classify atomic elements in text into predefined categories such as the names of persons, organizations, locations, expressions of times, quantities, monetary values, percentages, etc.” (Wiki: http://en.wikipedia.org/wiki/Named-entity_recognition)
Tokenisation: splitting a text into words/tokens (smallest units) based on colons, spaces, dashes, comas… (lecture notes); “Tokenization is the process of breaking a stream of text up into words, phrases, symbols, or other meaningful elements called tokens.” (Wiki: http://en.wikipedia.org/wiki/Tokenisation)
Morphological Analysis: looking for a canonical form (e.g. suppresses = suppress; activation = active) (lecture 05, slide 9); “In linguistics, morphology is the identification, analysis and description of the structure of a given language […] such as words, affixes, parts of speech, intonation/stress, or implied context” (Wiki: http://en.wikipedia.org/wiki/Morphology_%28linguistics%29)
POS-tagging: Part-of-speech tagging, assigning PoS-tags (verb, noun, adjective…) to tokens (lecture 05, slide 11 plus lecture notes); “A simplified form of this is commonly taught to school-age children, in the identification of words as nouns, verbs, adjectives, adverbs, etc.” (that seems to be our level: school-age children; from Wiki: http://en.wikipedia.org/wiki/Pos-

tagging)

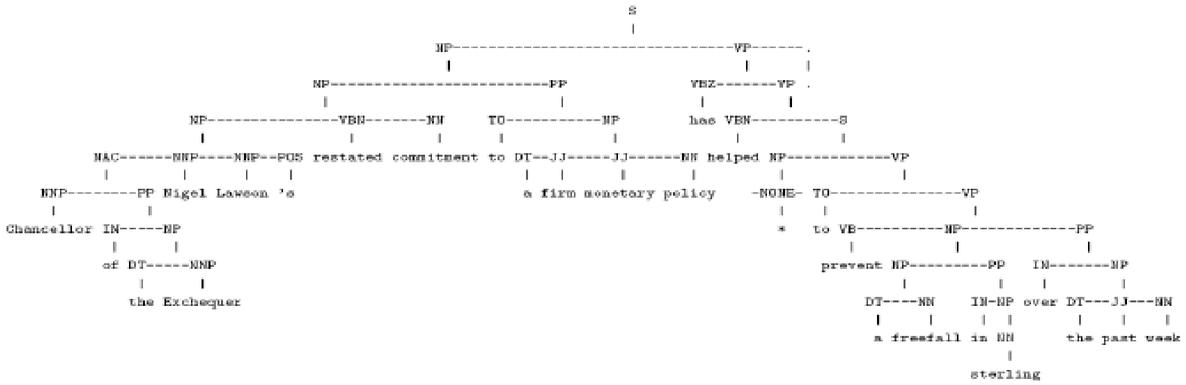
Syntax parsing: grouping words/chunks into phrases or clauses (lecture 05, slide 12)

Phrase Chunking: grouping of words in phrases (NPs, PPs. VPs etc.); use of PoS-tags, mainly NP-chunking (lecture 05, slide 11); Phrase chunking is a natural language process that separates and segments a sentence into its subconstituents, such as noun, verb, and prepositional phrases. (Wiki: http://en.wikipedia.org/wiki/Phrase_chunking)

69. What is the treebank corpus, which information is stored there?

The Treebank corpus contains (manually) annotated texts. The domain knowledge is necessary, it contains a clearly defined annotation guideline and in the ideal case at least three annotators should be involved. (lecture 05, slide 12)

It is a phrase structure tree with more than a hundred types of phrase labels. The phrase structure rules are relatively flat. It represents the phrase chunking of a sentence. (lecture 11, slide 12+14)


S
NP VP .
NP PP VBZ VP .
NP VBN NN TO NP has VBN S
NAC NNP NNP POS restated commitment to DT JJ JJ NN helped NP VP
NNP PP Nigel Lawson 's a firm monetary policy -NONE- TO VP
Chancellor IN NP * to VB NP PP
of DT NNP prevent NP PP IN NP
the Exchequer DT NN IN NP over DT JJ NN
a freefall in NN the past week
sterling

"A treebank or parsed corpus is a text corpus in which each sentence has been parsed, i.e. annotated with syntactic structure. Syntactic structure is commonly represented as a tree structure, hence the name Treebank." (Wiki: http://en.wikipedia.org/wiki/Treebank)

70. Explain the difficulties for NLP in general.

many words, many phenomena -> many rules; irregularities/exceptions (lecture 11, slide 2); sentence separation, word tokenization, verb forms, word frequencies and zipf's law (given some text corpus the frequency of any word is inversely proportional to its rank in the frequency table; lecture 11, slide 3), ambiguities (e.g. books can be a verb or noun) (lecture 11, slide 4 plus lecture notes)

71. Explain the difficulties for gene and protein name recognition.
72. Explain the difficulties for annotation.
73. Explain the two different approaches (mention recognition versus normalisation) for named entity recognition
74. What are advantages/disadvantages for using rule based/machine learning based systems?
75. What kind of information/features could be used for the development of a gene/ protein named entity approach?
76. Please introduce the Chomsky hierarchy and give the definitions for the different types
77. Which kind of grammar do you need to construct the following terms:
    a. (a*b*c*)*
    b. $a^n b^n$
    c. Write done the rules you need for the grammar
    d. Apply your rules for the following terms: abbca, (first rule set)
    e. Apply your rules for the following terms: aaaabbbb, (first rule set)
    f. Apply your rules for the following terms: aaaabbb (second rule set)
78. Define a probabilistic context free grammar
79. Develop a probabilistic context free grammar for the following training set; give the probability for every rule. Training set: [5 points]
    I. "The children eat chips with fingers."

II. “The children eat chips with ketchup.”

a. Define rules for the given training set. Use the following Nonterminals: NP,VP,PP, N, V, IN
b. Give the probability for every rule.
c. Build a tree for both sentences

80. Develop rules for the recognition of the following sentence:
calreticulin interacts with Glut-1
use the following set of nonterminals {S, VP, NP, PP, V, P, Protein}
81. Two annotators i.e. *annotator_A* and *annotator_B* are given the task to annotate a small piece of text with two classes *disease* and *chemical.* The text is mentioned below: [8 points]
“The study aim was to investigate the effect of salicylic acid peels for the treatment of post inflammatory hyperpigmentation. METHODS: Ten subjects with Fitzpatrick skin phototypes IV to VI were randomized to receive two 20% acidic peels to half of the face.”
The annotations are as follows:
annotator_A: salicylic acid (*chemical*), post inflammatory hyperpigmentation (*disease*)
annotator_B: salicylic acid (*chemical*), hyperpigmentation (*disease*), acidic (*chemical*)
Calculate the inter-annotator agreement (kappa) between the two annotators when Standard English punctuations and whitespaces are used during tokenization. Report the level of the agreement.

Application of Text Mining

82. What differences exist between random networks and scale-free networks?
83. Provide examples of Biomedical free-text resources and shortly describe what they contain.
84. Provide some examples of patent search scenarios.
85. What are the problems encountered when mining the full-text literature?
86. In what forms are the information about a *chemical* represented in full-text literature.
87. Provide examples of Biomedical and Chemistry databases and shortly describe what they contain.
88. List the document processing technologies that can be potentially helpful in mining the chemistry or pharma-based literature.
89. Give the general advantages and disadvantages of Name2Structure (N2S) conversion systems. Name few examples of N2S conversion tools.
90. Give the general advantages and disadvantages of Image2Structure (I2S) conversion systems. Name few examples of I2S conversion tools.
91. Provide examples and shortly describe the capabilities of different text search and visualization tools applicable for biomedical or chemistry domain.

Ontology-driven knowledge discovery

92. What is an ontology and why we need to develop ontologies?
93. Compare ontology engineering versus object-oriented modeling.
94. Draw the ontology life cycle and explain why ontology building is an iterative process?
95. What are the major steps in designing an ontology?
96. Define three modes of ontology development and explain which mode is preferred.
97. Explain why different ontologies should be orthogonal?
98. Sketch an ontological schema for the following concepts:
protein, ligand, protein-ligand interaction, interaction type, interaction descriptor, ligand activity, ligand conformer, protein type, protein fragment, receptor, enzyme.
99. What BFO (Basic Formal Ontology) represents and why it is used?
100. Describe OBO Foundry and its Relation Ontology.

101. Describe the process of ontology-driven knowledge discovery.