business data science
GRADUATE PROGRAM | Erasmus | VU

COURSE MANUAL
NATURAL LANGUAGE PROCESSING

Erasmus University Rotterdam, University of Amsterdam, Vrije Universiteit Amsterdam

Course Code	
Academic year	2019-2020
Period	May 4 – July 10
Credits	4
Recommended knowledge	Programming Basics, Mathematics, Statistics, Econometrics I, Supervised and Unsupervised Machine Learning
Required knowledge	Linear algebra, Regression, Machine Learning (e.g., classification, random forests, support vector machines)

Table of Contents

1. COURSE COORDINATOR AND LECTURERS

Coordinator/Lecturer: prof. dr. Bas Donkers (EUR)
Email: donkers@ese.eur.nl

Short bio: Bas Donkers is a professor of marketing research at the Erasmus School of Economics. His research examines consumer decision-making from a behavioural perspective and relies on the use of advanced quantitative analyses as well as various advanced market research techniques to establish new and ground breaking insights in the field. He has published articles in the leading journals in the field including the Journal of Marketing Research and Marketing Science.

business data science
GRADUATE PROGRAM | VU

Lecturer: dr. Meike Morren (VU)
Email: meike.morren@vu.nl
Short bio: Meike Morren is an Assistant Professor in Marketing at VU since 2012. She holds a PhD in Method and Statistics from Tilburg University. Her interests are sustainability, survey data quality, and text analysis. Her current projects involve NLP on restaurant reviews.

Teaching Assistant: see CANVAS
Email: see CANVAS

2. COURSE CONTENT

Natural Language Processing (NLP) comprises statistical and machine learning tools for automatically analysing text data to derive useful insights from it. Vast amounts of information are stored in this form, and hence NLP has become one of the essential technologies of the big data age. In this course, core concepts and techniques from the area will be studied, with a focus on methods that are popular in business applications. These include n-gram models, word vectors, sentiment analysis and topic modelling.

This course offers students a theoretically informed understanding of NLP. It aims at broadening the knowledge of the methods involved in NLP, as well as a hands-on experience with the steps that need to be taken in a NLP project. We focus on three aspects:

a) to create deep(er) understanding of the main methods in NLP (n-gram, lexicon approach, word2vec and other advanced machine learning methods);
b) to obtain an experience to scrape and clean the data yourself;
c) to apply this knowledge and experience in a group assignment which gives you the possibility to show your creativity.

By the end of this course, you will be able to analyse and evaluate NLP approaches. Moreover, you will apply this knowledge and skills in a real-life setting, enabling you to translate and apply theoretical knowledge into practice.

Topics covered:

1. Information theory, regular expressions and scraping (tokenization, stemming, lemmatization, parsing).
2. Word vectors and dimension reduction based on bag of words (n-grams, PCA)
3. Sentiment analysis (lexicon-based vs model-based)
4. Sentence completion (hidden Markov model, GPT and BERT)
5. Topic models (LDA) and word embeddings (GloVe, Word2Vec)

3. LEARNING OBJECTIVES

By the end of the course students will be able to:

- Understand the fundamentals of natural language processing including different ways of representing text data for statistical analysis,
- discuss and apply different sentiment analysis and topic modelling techniques,
- program selected algorithms involved in these methods, and
- be acquainted with NLP research areas.

KNOWLEDGE AND UNDERSTANDING	Demonstrate knowledge of the Natural Language processing, from data cleaning to advanced machine learning approaches, including: Information theory, regular expressions and scraping (tokenization, stemming, lemmatization, parsing). Word vectors and dimension reduction based on bag of words (n-grams, PCA) Sentiment analysis (lexicon-based vs model-based) Sentence completion (hidden Markov model, GPT and BERT) Topic models (LDA) and word embeddings (GloVe, Word2Vec)
APPLICATION OF KNOWLEDGE	Apply this theoretical knowledge and experience in group assignments. Analyze and evaluate NLP approaches.
COMMUNICATION	Written and oral presentation of results within the research project.
LEARNING SKILLS	Work on the research project in a team.

4. STUDY MATERIAL

The following list of mandatory readings (presented in alphabetical order) are considered essential for your learning experience. These articles are also part of the exam material. Changes in the reading list will be communicated on CANVAS.

Books:

- Jurafsky, D., & Martin, J. H. (2014). Speech and language processing (Vol. 3). London: Pearson.
- Manning, C. D., Manning, C. D., & Schütze, H. (1999). Foundations of statistical natural language processing. MIT press.

Selected papers, including:

- Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. Journal of machine learning research, 3(Jan), 993-1022.
- Mikolov, T.,Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In Advances in neural information processing systems (pp. 3111-3119).
- Hu, M., & Liu, B. (2004, July). Mining opinion features in customer reviews. In AAAI (Vol. 4, No. 4, pp. 755-760).
- Pennington, J., Socher, R., & Manning, C. (2014). Glove: Global vectors for word representation. In Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP) (pp. 1532-1543).
- Radford, Narasimhan, Salimans and Sutskever (2018), Improving Language Understanding by Generative Pre-Training, preprint.
- Devlin, Jacob, Ming-Wei Chang, Kenton Lee and Kristina Toutanova (2019) BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, preprint.

5. FORM OF TUITION

The lectures aim at stimulating your academic skills, and providing you with new knowledge. In this course, lectures are accompanied by tutorials and computer lab sessions. The information provided in the lecture is essential for the assignments and discussions during those sessions. We expect students to come to the lectures well prepared and to participate in the interaction.

The tutorials aim at practicing the theory using exercises and allowing students to ask for additional explanation for those parts of the material perceived as more difficult.

The computer lab sessions aim at making the material come alive and train students in how the methods learnt in class can actually be applied to data. The lab sessions are meant to work on the assignments, such that you automatically keep up with the material.

Class participation is a part of the course that is important, asking questions, contributing to answers and to the general learning atmosphere.

6. ASSESSMENT

Your overall course grade is composed based on different components that are presented in the assessment overview. You need a minimum overall grade of 5.50 to pass the course.

The final grade is rounded to the nearest multiple of .0 or .5, with the following exceptions: any grade between 5.0 and 5.5 is rounded to a 5; a 5.5 is rounded to a 6; a 0.5 does not exist. Grades for homework or midterm examinations do not need to be rounded.

Format	% grade	Knowledge and Understanding	Application of knowledge	Communication	Learning Skills
Exam	70%	X			
Quizzes	15%	X			
Assignments	15%*	X	X	X	X *Assignments are weighted by the student individual participation to the team.

Group Assignment - group assessment
During the first tutorial session, you will be assigned to a group of 3 students. The group assignments are designed to challenge you in various ways. They bring together different materials you study and practice during the lectures and tutorials, illustrating (some of) the principles learned in the course work and their applications (see **Appendix A**).

Group Participation - individual assessment
To assess your participation in the team we use an online peer evaluation system. You will assess both your teammates and yourself using a questionnaire (see **Appendix B**). This will result in a weighting factor that reflects the amount of effort and skills you have put in the group assignment. It consists of the average of the scores that the others give you on your role in the team and the assignment. From the early start of the project you know upon

which social skills you and your peers will evaluate each other. This means you can pay extra attention to these competences. These are the categories that you will focus on when evaluating your team members and yourself: Contributing to Work; Interacting with Teammates; Keeping Team on Track; Working Quality; Having Knowledge/Skills; and Team Satisfaction.

Class Participation - individual assessment
During the tutorials, you will be required to discuss, present, exchange information with fellow students and debate in different group settings. Hence, your presence is valued and essential in order for you to accomplish the objectives of this course (see **Appendix C** for the Assessment criteria).

Quizzes - individual assessment
The quizzes take place in week 2,4 and 6 of the course, during the tutorial on Thursdays. It will involve open-ended questions, which are testing knowledge, insight and application. The open-ended question templates are provided in **Appendix D**.

Written exam - individual assessment
The final exam takes place in week 8 of the course. The students will receive the exam on Monday, and are asked to hand it in on Thursday. It will involve open-ended questions, which are testing knowledge, insight and application. During the last week of the course, we will practice a few exam questions. The open-ended question templates are provided in **Appendix D**.

7. DETAILED COURSE SCHEDULE

Please check Canvas for an up-to-date schedule, reading material and assignments.

Week	Date	Time	Format	Theme/Topics	Preparation
1	Mon		Lecture 1 (2hrs)	**Scrape text, clean data**	*Mandatory readings:* ● Grammar ● Text structure
				Setting the Scene	Managing expectations & Project Kick-Off ● Team formation ● Selection of your website ● Explain case report ● Explain assignment 1 ● Explain what we expect
	Tue		Individual study and team work		
	Wed		Individual study and team work		
	Thu		Tutorial 1 (1hr)	**Project - Assignment 1:** Clean raw dataset	Work on assignment 1
	Fri		Individual study and team work		
2	Mon		Lecture 2 (1hr)	**Describe text using plots**	*Mandatory readings:* ● Ngrams ● Grammar ● Plots/PCA
			Tutorial 2 (1hr)	**Project – Assignment 1:** Make plots	Work on assignment 1 (see section 1)
	Tue		Individual study and team work		
	Wed		Individual study and team work		
	Thu		Tutorial 3 (1hr)	**Project – Assignment 1:** Write up results	Work on assignment 1 (see section 1) Quiz 1
	Fri	08.00	Submission Deadline	**Presentation**	Submit your presentation before 08.00 via canvas (see section 2)

Week	Date	Time	Format	Theme/Topics	Preparation
3	Mon		Lecture 3 (1hr)	**Sentiment analysis**	*Mandatory readings:* ● Sentiment analysis ● Dictionary/lexicon approach ● Word vectors / embeddings
			Tutorial 4 (1hr)		Presentations assignment 1 Explain assignment 2 (sentiment analysis, see section 3)
		17.00	Submission Deadline	Assignment 1	Submit your assignment before 17.00 via canvas
	Tue		Individual study and team work		
	Wed		Individual study and team work		
	Thu		Tutorial 5 (1hr)	**Project – Assignment 2:** Develop research question, clean data	Work on assignment 2
	Fri		Individual study and team work		
4	Mon		Lecture 4 (2hrs)	**Machine learning**	*Mandatory readings:* ● Prediction in machine learning ● Regression vs cluster approaches in machine learning ● Supervised vs unsupervised
	Tue		Individual study and team work		
	Wed		Individual study and team work		
	Thu		Tutorial 6 (1 hr)	**Q & A**	Team appointment (see section 4) Quiz 2
	Fri	08.00	Submission Deadline	**Presentation**	Submit your presentation before 08.00 via canvas (see section 2)
5	Mon		Lecture 5 (2hrs)	**Sentence completion**	*Mandatory readings:* ● Grammar / sentence structure (open AI) ● Hidden Markov Models ● Decision tree Presentation assignment 2 Explain assignment 3 (sentence completion, See section 5)
		17.00	Submission Deadline	**Assignment 2**	Submit your assignment before 17.00 via canvas
	Tue		Individual study and team work		
	Wed		Individual study and team work		
	Thu		Tutorial 7 (1hr)		Work on assignment 3
	Fri		Individual study and team work		
6	Mon		Lecture 6 (2hrs)	**Latent Dirichlet Allocation**	*Mandatory readings:* ● LDA ● Word2vec ● Elmo?
	Tue		Individual study and team work		
	Wed		Individual study and team work		
	Thu		Tutorial 8 (1hr)	**Q&A**	Team appointment

Week	Date	Time	Format	Theme/Topics	Preparation
					Quiz 3
	Fri	17.00	Submission Deadline	**Assignment 3**	Submit your assignment before 08.00 via canvas
7	Mon-Fri		Individual study and team work		
8	Mon-Wed		Individual study and team work		
	Thu	08.00	Submission take home exam		
	Fri	08.00	Submission Deadline	**Presentation**	Submit your presentation before 08.00 via canvas (see section 8)
					Final presentations

8. ASSIGNMENT INFORMATION

Section 1 Group assignment 1
In this assignment, you need to choose a website from which you will scrape reviews or comments. Formulate a research question that can be answered using these data. An example could be: what themes come up in the reviews/comments? You will clean and preprocess the text so that you can start to analyze the data. The second part of the first assignment is to plot and describe the data in an appealing way to the reader. The results should be reported in a concise manner in essay format (max 5 pages).

Section 2 Presentation assignment 1
You need to present your findings of assignment 1. Make a presentation for maximally 10 minutes, and prepare your presentation as a pitch. Thus, do not explain everything but highlight your most interesting findings, and explain how you achieved these findings, and what your main conclusions are. Usually, presenting one slide requires 2-3 minutes.

Section 3 Group assignment 2
In this assignment you will be given a set of reviews that still need to be processed so that you can start to analyze them. Your aim will be to get the sentiment from the reviews. The most simplistic would be positive, neutral and negative, but you are free to also make distinctions between very positive (negative) and moderate positive (negative). You make a choice and apply a method that has been discussed during the lecture. You need to include arguments in the report (essay format, max 5 pages).

Section 4 Team Q&A
Prepare yourself with questions you would like to ask the teacher. Per team a limited amount of time can be reserved so focus on the most important issues.

Section 5 Group assignment 3
In this assignment you will use word embeddings (GloVe or Word2Vec) and determine the optimal way to aggregate the word embeddings across documents in a corpus of your choice such that a specific feature of a document is predicted best. You will report on the various approaches studied and provide interpretation on the (most important) relationships that you find based on your most preferred aggregation method report (essay format, max 5 pages).

Section 8 Final presentation

In this final presentation, you present your three assignments, and discuss how they are connected. You focus on your learning curve, and how the theory has guided your choices in analysis and preparation.

business
data science
GRADUATE PROGRAM | Erasmus VU

APPENDIX A – GROUP ASSIGNMENTS ASSESSMENT A

Criterion	5 or lower	6	7	8	9 or 10
Knowledge/Application of knowledge **Weight 50%**	The assignment does not address the question(s). It contains evident logical errors or omissions. The answer is too simple or too limited for program or study load.	The assignment addresses sufficiently the question(s). It does not reflect all material covered, it includes considerable simplifications or shortcuts. Minimum level of adequacy for study load.	The assignment clearly addresses the question(s). It shows good knowledge of the material covered, awareness and own reflection of the important aspects of the material covered.	Well-considered and well-explained answer. Clear evidence of very good knowledge and understanding of the material covered, without being exceptional.	The assignment addresses fully the question(s). The students fully master the material covered in the course.
Application of knowledge **Weight 20%**	Insufficient ability to apply the knowledge covered. Major flaws.	Sufficient ability to apply the knowledge covered. Despite several flaws in the assignment, the outcome is satisfactory.	Good ability to apply the knowledge covered, with a few oversights.	Very good ability to apply the knowledge covered in the course.	Excellent ability to apply the knowledge covered in the program to different settings/data.
Written communication **Weight 5%**	Unstructured text. Fails to convey the key message of the thesis and/or to address questions. The text does not meet the academic editorial standards.	The text is somewhat unstructured and unclear. The text barely passes the academic editorial standards, as more polishing work is needed.	Overall well written, with occasional typos, or inaccuracies. The text passes the academic editorial standards, although the writing style is mechanical.	Structured text. The text is clear and concise, but here and there more (or less) details could improve the readability. Tables and Figures are self-explicatory and timely introduced in the text. The text meets the academic editorial standards, although the writing style is a bit mechanical at times.	Structured, coherent and polished text. Excellent writing style. The text is accurate, clear and concise, with the right level of detail. Tables and Figures are self-explicatory and timely introduced in the text. The text meets the academic editorial standards.
Presentation **Weight 5%**	Unstructured presentation. Fails to convey the key message of the thesis and/or to address questions.	Structured oral presentation. Answers and comments by the audience are not always adequately addressed.	Structured oral presentation. Answers and comments by the audience are adequately addressed most of the times.	Good oral presentation: well structured, right level of detail. Gives accurate and to the point response to comments and questions	Excellent oral presentation: coherent, right level of detail, lively. Gives accurate and to the point response to comments and questions
TOTAL					

business
data science
GRADUATE PROGRAM | Erasmus VU

APPENDIX B – GROUP PARTICIPATION ASSESSMENT

Criterion	Poor	Fair	Good	Very Good	Excellent
Contributing to the team's work **Weight 20%**	Does not do a fair share of the team's work. Delivers sloppy or incomplete work. Misses deadlines. Is late, unprepared, or absent for team meetings. Does not assist teammates. Quits if the work becomes difficult.	Demonstrates behaviors described immediately left and right.	Completes a fair share of the team's work with acceptable quality. Keeps commitments and completes assignments on time. Helps teammates who are having difficulty when it is easy or important.	Demonstrates behaviors described immediately left and right.	Does more or higher-quality work than expected. Makes important contributions that improve the team's work. Helps teammates who are having difficulty completing their work.
Interacting with teammates **Weight 20%**	Interrupts, ignores, bosses, or makes fun of teammates. Takes actions that affect teammates without their input. Does not share information. Complains, makes excuses, or does not interact with teammates. Is defensive. Will not accept help or advice from teammates.	Demonstrates behaviors described immediately left and right.	Listens to teammates and respects their contributions. Communicates clearly. Shares information with teammates. Participates fully in team activities. Respects and responds to feedback from teammates.	Demonstrates behaviors described immediately left and right.	Asks for and shows an interest in teammates' ideas and contributions. Makes sure teammates stay informed and understand each other. Provides encouragement or enthusiasm to the team. Asks teammates for feedback and uses their suggestions to improve.
Keeping the team on track **Weight 20%**	Is unaware of whether the team is meeting its goals. Does not pay attention to teammates' progress. Avoids discussing team problems, even when they are obvious.	Demonstrates behaviors described immediately left and right.	Notices changes that influence the team's success. Knows what everyone on the team should be doing and notices problems. Alerts teammates or suggests solutions when the team's success is threatened.	Demonstrates behaviors described immediately left and right.	Watches conditions affecting the team and monitors the team's progress. Makes sure that teammates are making appropriate progress. Gives teammates specific, timely, and constructive feedback.
Expected quality **Weight 20%**	Satisfied even if the team does not meet assigned standards. Wants the team to avoid work, even if it hurts the team. Doubts that the team can meet its requirements.	Demonstrates behaviors described immediately left and right.	Encourages the team to do good work that meets all requirements. Wants the team to perform well enough to earn all available rewards. Believes that the team can fully meet its responsibilities.	Demonstrates behaviors described immediately left and right.	Motivates the team to do excellent work. Cares that the team does outstanding work, even if there is no additional reward. Believes that the team can do excellent work.
Having related knowledge, skills, and abilities **Weight 20%**	Missing basic qualifications needed to be a member of the team. Unable or unwilling to develop knowledge or skills to contribute to the team. Unable to perform any of the duties of other team members.	Demonstrates behaviors described immediately left and right.	Demonstrates sufficient knowledge, skills, and abilities to contribute to the team's work. Acquires knowledge or skills as needed to meet requirements. Able to perform some of the tasks normally done by other team members.	Demonstrates behaviors described immediately left and right.	Demonstrates the knowledge, skills, and abilities to do excellent work. Acquires new knowledge or skills to improve the team's performance. Able to perform the role of any team member if necessary.
TOTAL					

business
data science
GRADUATE PROGRAM | Erasmus VU

APPENDIX C – CLASS PARTICIPATION ASSESSMENT

Criterion	Poor	Average	Good	Excellent
Level of engagement **Weight 33%**	Student never contributes to class by offering ideas and asking questions and/or has trouble staying on task during group project time.	Student rarely contributes to class by offering ideas and asking questions and/or works on group project only some of the allotted time.	Student proactively contributes to class by offering ideas and/or asks questions once per class and/or works on group project for most of the allotted time.	Student proactively contributes to class by offering ideas and/or asks questions more than once per class and/or works consistently on group project the entire time.
Quality of comments **Weight 33%**	Comments are uninformative, lacking in appropriate terminology. Heavy reliance on opinion and personal taste, e.g., "I love it", "I hate it", "It's bad" etc.	Comments are sometimes constructive, with occasional signs of insight. Student does not use appropriate terminology; comments not always relevant to the discussion.	Comments mostly insightful and constructive; mostly uses appropriate terminology. Occasionally comments are too general or not relevant to the discussion.	Comments always insightful and constructive; uses appropriate terminology. Comments balanced between general impressions, opinions and specific, thoughtful criticisms or contributions.
Listening skills **Weight 33%** **TOTAL**	Student does not listen when others talk, both in groups and in class. Student often interrupts when others speak. Student displays disruptive behavior during class.	Student does not really listen when others talk, both in groups and in class.	Student listens when others talk, both in groups and in class.	Student listens when others talk, both in groups and in class. Student incorporates or builds off of the ideas of others.

APPENDIX D – EXAMPLES OF QUIZ AND EXAM QUESTIONS

Quiz question

Latent Dirichlet Allocation [LDA] is one of the most popular techniques for detecting topics in texts.

a. Explain why LDA is often called a soft-clustering method, also carefully explain *what* is clustered and how cluster membership is measured.
b. Also explain why LDA is called a generative model. Furthermore, explain how topics are involved in this generative process.
c. Can LDA really be used to generate meaningful text? Explain why/why not.
d. Describe an example on how LDA can be used in a business context. Highlight why the bag of words assumption is appropriate in this context.

Exam question

A researcher uses a binary Generalized Linear Model with a logit link to analyze 5,000 reviews on mobile phones. The researcher studies the relation between a positive/negative product rating and the content of the review text. The review text is summarized in a number of features:

- Number of words in the review
- Indicator whether a certain stemmed word appears in a given review.
- Indicator whether a certain stemmed bi-gram appears in a given review.

At first the researcher applies a model using the number of words, a few words, and a few bi-grams as explanatory variables. The dependent variable equals 1 in case the review is positive. The results are as follows:

Variable	Coefficient
Intercept	-0.260
Number of words	0.012
“batteri”	-0.101
“screen”	0.051
“di” (derived from dies)	-0.142
“heavi”	-0.078
“big”	-0.101
“slow”	-0.210
“batteri life”	0.145
“larg screen”	0.124

You can assume that all estimated effects are significant.

a. Clearly explain the impact of the use of the bi-gram “batteri life” in a review on the rating. Be very precise.
b. Suppose that a specific review has probability 0.5 of being positive. How does the expected rating change in case the author of the review decides to use more stop words in his review?
c. Calculate the probability of a positive review for the following text: “This phone has a beautiful large screen but the battery dies quickly.”
d. The researcher has also obtained word embeddings using the GloVe method. How can these embeddings be used in the context of predicting the product rating?
e. For the final model, the researcher considers to include many explanatory variables. He has identified 1500 words, 750 bi-grams, 50 variables based on GloVe and 10 sentiment related variables. Explain the overfitting problem and describe how this problem can be circumvented by the researcher.