**שם הקורס בעברית**: היבטים מעשיים בכריית תוכן אינטרנטי לאפליקציות עסקיות

**שם הקורס באנגלית**: Practical Web Mining and Modeling for Business Applications

מספר הקורס: 372.1.4991

מבנה הקורס: הרצאה: 3 , תרגיל 0, מעבדה 0, סה"כ: 3 ש"ש

נקודות זכות: 3

תאור הקורס:

הקורס מספק ידע מעשי עדכני בשיטות כרית, עיבוד ומידול מידע אינטרנטי לצורך אפליקציות חיזוי עסקיות.

מטרות הקורס:

הנושאים העיקרים כוללים בניית "כורים", סינון וקטלוג תוכן, שיוך רגשי, חיזוי סדרות זמן ומגמות בעזרת שירותים כגון Twitter streaming , Google Trends וחדשות.

תנאי קדם:

מבוא להסתברות וסטטיסטיקה, תכנות מתקדם

דרישות הקורס והרכב הציון:

40% מבחן, 60% שיעורי בית ופרויקט מסכם

רשימת קריאה (בבליוגרפיה):

[WSP] Web scraping with Python. O’Reilly Media. Ryan M. (2015)

[MSW] Mining the Social Web: Analyzing Data from Facebook,

Twitter, LinkedIn, and Other Social Media Sites. O’Reilly Media (2011). Russell, M.

[INLP][Introduction to Natural Language Processing](http://www.nltk.org/book), by Steven Bird, Ewan Klein and Edward Loper, 2009, distributed on the NLTK site

[DMML] Data Mining: Practical Machine Learning Tools and Techniques, Ian H. Witten, Eibe Frank and Mark A. Hall, Third Edition, Morgan Kaufmann, 2011.

[TML] [David Blei's Lecture on LDA](http://videolectures.net/mlss09uk_blei_tm/) Sept 2009, Part 1 (1h30) and Part 2 (1h30)

[GCGT] [Predicting the present with Google Correlate and Google Trends](http://social-dynamics.org/predicting-the-present-with-google-correlate-and-google-trends/)

[SNTA] [Text Mining Analytics Online Course (Coursera) - Week 5](https://www.coursera.org/learn/text-mining-analytics/home/welcome)

[Semantic Tagging Using Topic Models Exploiting Wikipedia Category Network](https://www.researchgate.net/profile/Mehdi_Allahyari/publication/301701156_Semantic_Tagging_Using_Topic_Models_Exploiting_Wikipedia_Category_Network/links/5723c77f08aef9c00b811cbd/Semantic-Tagging-Using-Topic-Models-Exploiting-Wikipedia-Category-Network.pdf) (48 minutes)

[WE1]  [CS 224D: Deep Learning for NLP1 1 - Lecture Notes: Part I](http://cs224d.stanford.edu/lecture_notes/LectureNotes1.pdf) by Richard Socher, 2015 and the links of the first chapter in [Deep Learning for NLP](http://cs224d.stanford.edu/syllabus.html) course.

[WE2] [Efficient Estimation of Word Representations in Vector Space](http://arxiv.org/pdf/1301.3781.pdf) from Mikolov et al (2013)

[WE3] Pennington, J., Socher, R. and Manning, C., 2014. Glove: [Global vectors for word representation](http://www.aclweb.org/anthology/D14-1162). In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)* (pp. 1532-1543).

[RNN] [Natural Language Understanding with Distributed Representation](http://arxiv.org/abs/1511.07916) by Kyunghyun Cho, Nov 2015 (chapters 4,5)

[TW1] [Stock prediction using Twitter sentiment analysis (2012)](http://tomx.inf.elte.hu/twiki/pub/Tudas_Labor/2012Summer/GoelMittal-StockMarketPredictionUsingTwitterSentimentAnalysis.pdf)

[TW2] [Tweets and votes, a special relationship: the 2009 federal election in germany](http://dl.acm.org/citation.cfm?id=2508437) (2013)

[TW3] [Twitter sentiment analysis with deep convolutional neural networks (2015).](http://dl.acm.org/citation.cfm?id=2767830)

תכנית המפגשים:

|  |  |  |
| --- | --- | --- |
| **קריאה נדרשת** | **נושא השיעור** | **מס' השיעור** |
|  | Introduction | 1 |
| [WSP] | Building web crawlers | 2 |
| [INLP] (chapters 1-4,7) | Text mining, preprocessing and normalization | 3 |
| [DMML] | Classification and categorization machine learning models | 4 |
| [INLP] (chapters 5-6) | TFIDF and document classification/categorization | 5 |
| [SNTA] | Text sentiment analysis | 6 |
| [TML] | Topic modeling | 7 |
|  | Text classification lab + guest lecture ([hebhose.io](https://webhose.io/)) | 8 |
| [WE1], [WE2], [WE3] | Word embedding | 9 |
| [GCGT] | Predicting the present with Google Correlate and Google Trends | 10 |
| [RNN] | RNN and sequence classification | 11 |
| [MSW](chapters 1, 9) | Twitter text mining | 12 |
| [TW1], [TW2], [TW3] | Twitter for business applications | 13 |