NORTHWESTERN UNIVERSITY | SCHOOL OF CONTINUING STUDIES

PREDICT 453: Text Analytics Syllabus Spring 2013

Kenneth Stehlik-Barry, Ph.D.

Kenneth.stehlik-barry@northwestern.edu

Office Hours: By appointment

Course Description

This course will delve into business problems that exist in the query and analysis of data in an unstructured form only bond by linguistic structure of the data in its natural language. Topics include information or data retrieval, lexical analysis to understand word frequency distributions, pattern recognition, tagging or annotation application, and information extraction. A review of data mining will also be incorporated due to the importance of link and associational analysis and visualization.

Text

Weiss, S., Indurkhya, N. & Zhang, T (2010). *Fundamentals of Predictive Text Mining*. New York, NY: Springer.

[ISBN-13: 978-1849962254]

Software

IBM SPSS Modeler Version 15 with Text Analytics

Prerequisites

PREDICT 401 and PREDICT 410

Learning Goals

The goals of this course are to:

- Describe potential sources of unstructured data within organizations and external to organizations.
- Explain the ways in which various unstructured data sources could be used as part of a predictive analytics solution.
- Extract meaning from unstructured data.
- Perform sentiment analysis.
- Combine concepts found in unstructured data via text link analysis.
- Integrate unstructured data with structured data to build predictive models.

Evaluation

The student's final grade will be determined as follows:

- Document Samples: 25 pts.
- Working with XML Tags: 50 pts.
- RSS Feeds: 50 pts.
- Text Mining: 50 pts.
- Synonym Customization: 50 pts.
- Dictionary Customization: 50 pts.
- Sentiment Analysis: 50 pts.
- Text Link Analysis: 50 pts.
- Deployment Strategy: 25 pts.
- Discussion Board Participation: 100 pts.

Total Points: 500 pts.

Grading Scale

A = 93%–100% (465–500 pts.)

A- = 90%–92% (450–464 pts.)

B+ = 87%–89% (435–449 pts.)
B = 83%–86% (415–434 pts.)
B- = 80%–82% (400–414 pts.)
C+ = 77%–79% (385–399 pts.)
C = 73%–76% (365–384 pts.)
C- = 70%–72% (350–364 pts.)
F = 0%–69% (0–349 pts.)

Discussion Board Etiquette

The purpose of the discussion boards is to allow students to freely exchange ideas. It is imperative to remain respectful of all viewpoints and positions and, when necessary, agree to respectfully disagree. While active and frequent participation is encouraged, cluttering a discussion board with inappropriate, irrelevant, or insignificant material will not earn additional points and may result in receiving less than full credit. Frequency is not unimportant, but content of the message is paramount. Please remember to cite all sources—when relevant—in order to avoid plagiarism.

Proctored Assessment

There is no proctored assessment requirement for this class.

Attendance

This course will not meet at a particular time each week. All course goals, session learning objectives, and assessments are supported through classroom elements that can be accessed at any time. To measure class participation (or attendance), your participation in threaded discussion boards is required, graded, and paramount to your success in this class. Please note that any scheduled synchronous or "live" meetings are considered supplemental and optional. While your attendance is highly encouraged, it is not required and you will not be graded on your attendance or participation.

Late Work

Students must provide written notification of late work 24 hours in advance of the deadline. The grace period is 24 hours for those who provide the require advance late work notification. A maximum of two grace periods for the course is allowed without reduction of points. A 25% reduction is applied to the grade for every 24 hours late. No negative points are applied.

Academic Integrity at Northwestern

Students are required to comply with University regulations regarding academic integrity. If you are in doubt about what constitutes academic dishonesty, speak with your instructor or graduate coordinator before the assignment is due and/or examine the University Web site. Academic dishonesty includes, but is not limited to, cheating on an exam, obtaining an unfair advantage, and plagiarism (e.g., using material from readings without citing or copying another student's paper). Failure to maintain academic integrity will result in a grade sanction, possibly as severe as failing and being required to retake the course, and could lead to a suspension or expulsion from the program. Further penalties may apply. For more information, visit <www.scs.northwestern.edu/student/issues/academic_integrity.cfm>.

Plagiarism is one form of academic dishonesty. Students can familiarize themselves with the definition and examples of plagiarism, by visiting <www.northwestern.edu/uacc/plagiar.html>. A myriad of other sources can be found online.

Some assignments in this course may be required to be submitted through SafeAssign, a plagiarism detection and education tool. You can find an explanation of the tool at <http://wiki.safeassign.com/display/SAFE/How+Does+SafeAssign+Work>. In brief, SafeAssign compares the submitted assignment to millions of documents in large databases. It then generates a report showing the extent to which text within a paper is similar to pre-existing sources. The user can see how or whether the flagged text is appropriately cited. SafeAssign also returns a percentage score, indicating the percentage of the submitted paper that is similar or identical to pre-existing sources. High scores are not necessarily bad, nor do they necessarily indicate plagiarism, since the score does not take into account how or whether material is cited. If a paper consisted of one long quote that was cited appropriately, it would score 100%. This would not be plagiarism, due to the appropriate citation. However, submitting one long quote would probably be a poor paper. Low scores are not necessarily good, nor do they necessarily indicate a lack of plagiarism. If a 50-page paper contained all original material, except for one short quote that was not cited, it might score around 1%. But, not citing a quotation is still plagiarism.

SafeAssign includes an option in which the student can submit a paper and see the resultant report before submitting a final copy to the instructor. This ideally will help students better understand and avoid plagiarism.

Other Processes and Policies

Please refer to your SCS student handbook at <www.scs.northwestern.edu/grad/information/handbook.cfm> for additional course and program processes and policies.

Course Schedule

Important Note: Changes may occur to the syllabus at the instructor's discretion.
When changes are made, students will be notified via an announcement in Blackboard.

Session 1

Learning Objectives

After this session, the student will be able to:

- Identify relevant sources of unstructured data for various areas of interest.
- Describe the factors that determine the contribution these data sources likely bring to the predictive process.
- Describe the types of information about organizations that can be obtained by leveraging the information in publically available documents.
- Describe structured data fields that could be derived from unstructured data.

Course Content

Textbook Reading

Weiss, Indurkhya, Zhang, and Damerau, Chapter 1

Additional Reading

Jurafsky and Martin, *Speech and Language Processing*

Inmon and Nesavich, *Tapping into Unstructured Data*

Online Video

NOVA: Smartest Machine on Earth

Discussion Board

Each session you are required to participate in the session-specific discussion board forum. Your participation in both posting and responding to other students' comments is graded. For this session's discussion topic(s), visit the discussion board in Blackboard.

Assignments

None.

Sync Session

Tuesday, April 2, 2013 at 7:00–9:30 p.m. (central time)

Session 2

Learning Objectives

After this session, the student will be able to:

- Describe the value of data that can be extracted from unstructured data sources.
- Describe the types of unstructured data sources from various domains that could be used in conjunction with predictive analytics.
- Describe the appropriate XML tags to include from source data given a specific business situation.

Course Content

Data Sample Documents

U.S. Patent

Telecommunications, Inc.

SPSS Earnings

U.S. Bank Earnings

Webcast

U.S. Bank Earnings Call

XML Samples

SPSS Earnings

U.S. Bank Earnings Call

U.S. Bank Earnings Call Metadata

Discussion Board

Each session you are required to participate in the session-specific discussion board forum. Your participation in both posting and responding to other students' comments is graded. For this session's discussion topic(s), visit the discussion board in Blackboard.

Assignments

Document Samples is due at Sunday, April 14, 2013 at 11:55 p.m. (central time). For more information, click Assignments on the left navigation panel in Blackboard, and scroll to this assignment's item.

Working with XML Tags is due at Sunday, April 14, 2013 at 11:55 p.m. (central time). For more information, click Assignments on the left navigation panel in Blackboard, and scroll to this assignment's item.

Sync Session

None.

Session 3

Learning Objectives

After this session, the student will be able to:

- Read textual data using RSS and other web-based formats.
- Extract terms and concepts from textual data.
- Organize terms and concepts into categories.

Course Content

Textbook Reading

Weiss, Indurkhya, Zhang, and Damerau, Chapters 2 and 3

Additional Reading

Center for Disease Control, *E. Coli Outbreaks*

Pinker, *Words and Rules*

IBM, *User's Guide*, Chapter 1

Lecture Capture

Structured Fields Organization

Discussion Board

Each session you are required to participate in the session-specific discussion board forum. Your participation in both posting and responding to other students' comments is graded. For this session's discussion topic(s), visit the discussion board in Blackboard.

Assignments

RSS Feeds is due at Sunday, April 21, 2013 at 11:55 p.m. (central time). For more information, click Assignments on the left navigation panel in Blackboard, and scroll to this assignment's item.

Text Mining is due at Sunday, April 21, 2013 at 11:55 p.m. (central time). For more information, click Assignments on the left navigation panel in Blackboard, and scroll to this assignment's item.

Sync Session

None.

Session 4

Learning Objectives

After this session, the student will be able to:

- Relate categories to one another and to individual concepts to capture the meaning in text.
- Demonstrate how to incorporate synonyms and alias logic into standard extraction processes.
- Modify extraction dictionaries to address domain-specific needs and terminology.

Course Content

Additional Reading

Miner, Elder, Hill, Nisbet, Delen, and Fast, *Practical Text Mining and Statistical Analysis for Non-Structured Text Data Applications*, Chapter 4

IBM, *User's Guide*, Chapters 3 and 10

Discussion Board

Each session you are required to participate in the session-specific discussion board forum. Your participation in both posting and responding to other students' comments is graded. For this session's discussion topic(s), visit the discussion board in Blackboard.

Assignments

Synonym Customization is due Sunday, April 28, 2013 at 11:55 p.m. (central time). For more information, click Assignments on the left navigation panel in Blackboard, and scroll to this assignment's item.

Sync Session

None.

Session 5

Learning Objectives

After this session, the student will be able to:

- Compile custom extraction dictionaries from domain specific sources.
- Describe the challenges involved in gleaning positive and negative sentiment from unstructured data.
- Connect negative and positive sentiment to the related concepts and topics to obtain the underlying meaning.

Course Content

Additional Reading

Miner, Elder, Hill, Nisbet, Delen, and Fast, *Practical Text Mining and Statistical Analysis for Non-Structured Text Data Applications*, Chapter 5

Lecture Capture

Issues with Developing Effective Dictionaries

Discussion Board

Each session you are required to participate in the session-specific discussion board forum. Your participation in both posting and responding to other students' comments is graded. For this session's discussion topic(s), visit the discussion board in Blackboard.

Assignments

Dictionary Customization is due Sunday, May 5, 2013 at 11:55 p.m. (central time). For more information, click Assignments on the left navigation panel in Blackboard, and scroll to this assignment's item.

Sync Session

None.

Session 6

Learning Objectives

After this session, the student will be able to:

- Quantify the extent of negative and positive sentiments in unstructured data.
- Extract sentiment from open-ended comment fields in surveys.
- Connect sentiment expressed in survey comments to the topics being discussed in the corresponding text.

Course Content

Additional Reading

Pestian, Matykiewicz, Linn-Guest, South, Uzuner, *Sentiment Analysis of Suicide Notes*

Discussion Board

Each session you are required to participate in the session-specific discussion board forum. Your participation in both posting and responding to other students' comments is graded. For this session's discussion topic(s), visit the discussion board in Blackboard.

Assignments

None.

Sync Session

None.

Session 7

Learning Objectives

After this session, the student will be able to:

- Associate satisfaction and dissatisfaction expressed in textual comments to specific product or service attributes mentioned.
- Link sets of themes captured from unstructured data to more fully capture the relevant information contained in the text.
- Detect situations in which key patterns of linkage relevant to specific subject matter domains occur.

Course Content

Additional Reading

IBM, *User's Guide*, Chapters 12, 15, 16, and 17

Lecture Capture

Sentiment Analysis

Multimedia

Link Analysis

Discussion Board

Each session you are required to participate in the session-specific discussion board forum. Your participation in both posting and responding to other students' comments is graded. For this session's discussion topic(s), visit the discussion board in Blackboard.

Assignments

Sentiment Analysis is due Sunday, May 19, 2013 at 11:55 p.m. (central time). For more information, click Assignments on the left navigation panel in Blackboard, and scroll to this assignment's item.

Sync Session

None.

Session 8

Learning Objectives

After this session, the student will be able to:

- Describe the process for establishing connections to access required data from various sources on a regular basis.
- Apply key patterns of linkage to new documents.

Course Content

Additional Reading

IBM, *User's Guide*, Chapters 4 and 19 (pages 308–315)

Discussion Board

Each session you are required to participate in the session-specific discussion board forum. Your participation in both posting and responding to other students' comments is graded. For this session's discussion topic(s), visit the discussion board in Blackboard.

Assignments

Text Link Analysis is due Sunday, May 26, 2013 at 11:55 p.m. (central time). For more information, click Assignments on the left navigation panel in Blackboard, and scroll to this assignment's item.

Sync Session

None.

Session 9

Learning Objectives

After this session, the student will be able to:

- Set up a scoring process to apply previously built text analysis models to new documents and feeds.
- Discuss how the scores produced by the ongoing analysis of this unstructured data can be used to by an organization to improve interactions with customers/clients.
- Describe how changes over time in the information captured from the unstructured data can be presented.

Course Content

Additional Reading

Chapman, Clinton, Kerber, Khabaza, Reinartz, Shearer, and Wirth, *CRISP-DM 1.0: Step-by-Step Data Mining Guide*

Multimedia

Deploying Text Analytics into Operational Systems

Discussion Board

Each session you are required to participate in the session-specific discussion board forum. Your participation in both posting and responding to other students' comments is graded. For this session's discussion topic(s), visit the discussion board in Blackboard.

Assignments

Deployment Strategy is due Sunday, June 2, 2013 at 11:55 p.m. (central time). For more information, click Assignments on the left navigation panel in Blackboard, and scroll to this assignment's item.

Sync Session

None.

Session 10

Learning Objectives

After this session, the student will be able to:

- No new learning objectives.

Course Content

None.

Discussion Board

Each session you are required to participate in the session-specific discussion board forum. Your participation in both posting and responding to other students' comments is graded. For this session's discussion topic(s), visit the discussion board in Blackboard.

Assignments

None.

Sync Session

None.

