Proposal for a Text Analytics Minor and a Basic Certificate (attachments)

Jean Mark Gawron
Rob Malouf

Sept 21, 2015

1 Introduction

Data science is a fast growing new discipline that combines elements of machine learning, statistical analysis, visualization, computer science, and linguistics. One of the most important areas of this new field is text analytics, the analysis, visualization, and mining of unstructured free text. Although we have course offerings in computational linguistics, statistics, and computer science that all contain elements of a text analytics program, as well as a computational linguistics cvertificate, as yet there is no single program focusing on analytics and the special issue text analytics raise, and offering students a credential to help them find jobs in this specialized area.

2 Description

Two program proposals have been submitted on CurricUnet, a Text Analytics Minor and a Text Analytics Basic Certificate. The proposals are almost identical, except that the Minor includes 3 units of electives. The core courses in the Text Analytics Minor also serve as the core courses in a basic certificate. The two programs include foundation courses for language analysis and programming, some statistics background to provide a better foundation for understanding the machine learning methods that have become so central in text analytics, and a new capstone course (Ling 583) to ground that deeper understanding in practice. Our future plans include proposing an advanced certificate and a new graduate course that will be the capstone for that.

The combination of Ling 571 or 572, Ling 581, and Stat 550 (or an equivalent or more advanced course on Probability) will provide students with foundational skills, programming and the use of analytical tools appropriate for language. They will also prepare students for the more advanced treatment of state of the art text analysis methods to be given in Ling 583 (Statistical Methods in Text Analysis), a new course. The goal of this capstone course is to enable students to become critical and creative users of this new technology. For the minor, electives in Statistics, Computer Science and Linguistics enable students to deepen their knowledge of machine learning methods on text by applying them to other domains.

2.1 Differences between this program and existing programs

The Linguistics Department offers a certificate in Computational Linguistics, which has overlapping course content (Ling 571,572, 581), but includes basic linguistics courses as well (420,501), and does not include Stat 550 (or equivalent probability course) or Ling 583. This new certificate targets students with more background in mathematics and statistics, and covers more advanced material in Ling 583.

The Department of Statistics already offers coursework in core data mining techniques and statistical methodology, as pointed out in Dr. O'Sullivan and Dr. Levine's letter of support, but they point out that the area of text analytics is a significant specialty within the broad area of data science: "text mining and unstructured data analysis have become a critical component in the statistics and data science marketplace."

2.2 Why both a minor and a certificate

The minor and basic certificate will offer students two mutually exclusive options. No course can be used for both the Minor and the Certificate. The purpose of proposing the Certificate along with the minor is so that interested graduate students can use the coursework as a credential. Hopefully, this will make the courses both more attractive and more practical for them. In addition, basic certificates can also be made available through the College of Extended Studies (CES), enabling non-matriculated students to participate. Our basic computational linguistics certificate has attracted a number of open university students through the years, and we expect the text analytics certificate to do the same. Although we plan to offer an advanced certificate specifically targeting the needs of graduate students, we decided to prioritize making the certificate available for Grad students in various departments. If the interest is there, we plan to propose a related advanced certificate later, so that interested off-campus particpants can enroll.

The Minor will require a minimum of 15 units of coursework, and the certificate will require 12 units.

3 Required Courses

Both the certificate and the minor will require the following core courses:

Ling 571 or 572	Corpus Linguistics or Python Scripting	3
Stat 550 or Stat 551A or Stat 670A	Applied Probability Theory (with R)	3
Ling 581	Intro to Computational Linguistics	3
Ling 583	Statistical Methods in Text Analysis	3
Total		12

Ling 571 and 572 both offer introductions to programming, but with somewhat different emphases. Ling 572 is intended more as a general introduction to programming, using Python because it is good first programming language, allowing simple programs to be written in few lines. Ling 571 is geared more toward text processing, addressing some big data issues with large text data sets.

Ling 583 is a new course being proposed in conjunction with the minor, intended to be taken after one of the Statistics courses covering probability theory.

4 Electives for Minor

The Minor in addition will require one of the following electives.

	1 of the following		
	CS 550	Artificial Intelligence I	3
	Stat 520	Applied Multivariate Analysis	3
	Ling 551	Sociolinguistics	3
	Biol 568	Informatics	3

5 Catalog copy

Minor The Minor in Text Analytics consists of a minimum of 15 units to include Linguistics 571 or 572; Linguistics 581; Linguistics 583, and Statistics 550 or equivalent (Statistics 551A or Statistics 670A). In addition, students must complete 1 of the following courses Stat 520, CS 550, Biol 568, or Ling 551. Courses in the minor may not be counted toward the major. A minimum of six upper division units must be completed in residence at San Diego State University.

Certficate The Basic Certficate in Text Analytics consists of 12 units to include Linguistics 571 or 572; Linguistics 581; Linguistics 583, and Statistics 550 or equivalent (Statistics 551A or Statistics 670A).

6 Program learning outcomes

Upon satisfactory completion of the program, students will be able to:

1. Describe a variety of text analysis steps, such as parsing, part of speech tagging, text classification, text clustering, information extraction, topic modeling, and sentiment analysis, understand their interdependencies, and identify algorithms for each step.
2. Define the difference between sequential and non-sequential probabilistic models and generative and discriminative models, and identify examples of each.
3. Choose an appropriate algorithm and/or analytic tool for a variety of practical text analysis problems, as well as provide explanatory presentations of their results (such as graphical visualizations).
4. Perform analyses of a variety of types on actual text data by writing scripts in a programming language such as R or Python that apply tools appropriate for the data and the task.
5. Apply an understanding of the ways that linguistic information is structured to the analysis of natural language texts.

7 Student Learning Outcomes

1. Describe a variety of text analysis steps, such as parsing, part of speech tagging, text classification, text clustering, information extraction, topic modeling, and sentiment analysis, understand their interdependencies, and identify algorithms for each step.

   Activity Ling 581 uses a text (Jurafsky & Martin) covering the basic content. Students either write programs that execute a step (part of speech tagger) or trace the steps as the algorithm applies to example texts (parsing).

   Assessment Both written programs and traces are handed in as homework assignments, and graded for functionality, correctness, and economy. A correct implementation or an understanding of the information flow in an algorithm reflects an understanding of what an analysis step does.

2. Define the difference between sequential and non-sequential probabilistic models and generative and discriminative models, and identify examples of each.

   Activity Ling 583 will be a project-oriented course presenting students with analysis problems and requiring them to design solutions in a series of steps.

   Assessment Problems will be chosen so as to require a variety of approaches, and feedback given early in the design process, so as to illustrate the drawbacks of approaches poorly suited to the problem. Revised designs will be encouraged to incorporate lessons learned.

3. Choose an appropriate algorithm and/or analytic tool for a variety of practical text analysis problems, as well as provide explanatory presentations of their results (such as graphical visualizations).

   Activity Ling 581 covers sequential and non-sequential probability models, as exemplified in part of speech taggers or speech recognizer language models and Naive Bayes Classifiers. All are generative. Discriminative models are introduced when Maximum Entropy part of speech taggers are studied. Students are introduced to generative Naive Bayes classifiers in Ling 581, via the text book and lecture, and the equations are explicitly compared to discriminative linear models. They are asked to assess the effects of the differing assumptions of generative NB and discriminative log linear models in a homework assignment, referring to the equations. In another assignment, students draw dependency diagrams for their models, making explicit whether history is a factor.

   Assessment The assignment in which students compare and discuss results from an NB classifier and a log linear classifier is graded for validity, clarity, and the explicitness with which they discuss the consequences of the model assumptions. The dependency diagram assignment is graded for accuracy and completeness.

4. Perform analyses of a variety of types on actual text data by writing scripts in a programming language such as R or Python that apply tools appropriate for the data and the task.

   Activity Programming assignments are part of Ling 571, 572, 581, and 583.

   Assessment Programs are graded with increasing rigor as students progress from 571 or 572 to 581 to 583. Early on, functionality is less of an issue, because the rules of a language are being learned, and even small mistakes can be fatal. The emphasis early on is on teaching them how to diagnose mistakes; quizzes and programming assignments reflect this emphasis. Later, system design becomes the goal, with well-defined interfaces serving the needs of a particular problem; functionality is required, but the grading emphasis is on design quality, measured by appropriateness of the program for the problem being solved.

5. Apply an understanding of the ways that linguistic information is structured to the analysis of natural language texts.

Activity Using regular expressions and syntactically parsed outputs, students are asked to write linguistic patterns for the extraction of information from texts, and will compare structure-based and regular-expression based extraction features.

Assessment This assignment will be graded on how well the student is able to exploit the advantage of using syntactic structure to write patterns.

8 Program Comprehensive Assessment Plan

The capstone course in the program is Ling 583 and this is the strategic point at which to assess student achievement in various practical skills. This will be a practice-oriented class in which students will engage in programming projects addressing different kinds of problems. As indicated in the student learning outcomes section, assessment of these projects will be of several different kinds:

- Assessing the appropriateness of the components in a data-processing pipeline.
- Assessing the appropriateness of the algorithms used in the step
- Assessing the design quality of the programs used to solve problems
- Assessing the functionality and correctness of the programs written (do they do what they're supposed to)

Ling 581 povides strategic points at which to assess understanding of key theoretical concepts:

- Assessing the understanding of key theoretical concepts like sequential and discriminative models, with assignments requesting analysis of example problems.
- Assessing the understanding of how linguistic structure plays a role in organizing information in apparently ”unstructured” text with information extraction exercises comparing structure-based with other approaches.

We will produce a document at the end of the first two years compiling student scores on critical assignments and programmng projects in these two classes.

9 Supporting information

9.1 Societal and public needs assesement

(a) List other California State University campuses currently offering or projecting the proposed degree major program; list neighboring institutions, public and private, currently offering the proposed degree major program.

As far as we know this is the first CSU campus to offer a comparable minor or certificate. UCSD offers a certificate in text mining and UC Berkeley an online masters in Data Science. Stanford offers an M.S. in Data Science. The University of Washington offers certificates in Data Science, Machine Learning, and Natural Language Technology, as well as an M.S. in Computational Linguistics.

(b) Describe differences between the proposed program and programs listed in (a)

There are no CSU programs to compare this one to. The UC San Diego certificate is close in concept and depth (4 courses, plus one elective), but it does not target specific issues of text analysis. Of the University of Washington offerings, the certificate in Natural Language Technology is the closest in subject matter and scope. The Minor we propose differs, however, in building in more mathematical background.

The UC Berkeley and Stanford programs are graduate degrees assuming their incoming students have significant technical background (UCB requires an undergrad linear algebra class). As such they cover both a broader set of topics (data mining with nontext data) and they cover topics in more depth.

(c) List other curricula currently offered by the campus that are closely related to the proposed program.

The Linguistics Department offers a certificate in Computational Linguistics, which has overlapping course content (Ling 571, 572, 581), but includes basic linguistics courses as well (420, 501), and does not include Stat 550 (or any equivalent) or Ling 583. This new certificate targets students with more background in mathematics and statistics, and covers more advanced material in Ling 583.

The Department of Statistics already offers coursework in core data mining techniques and statistical methodology, as pointed out in Dr. O'Sullivan and Dr. Levine's letter of support , but they point out that the area of text analytics is a significant specialty within the broad area of data science: "Text mining and unstructured data analysis have become a critical component in the statistics and data science marketplace." One of the Statistics students who took 572 the first time it was taught was working as an intern for Nordstroms, and they were interested in having him doing some analysis of Twitter data related to their products. His Statistics skills were crucial, but many of the questions they had required content analysis, not just sentiment analysis, but answers to why questions about the sentiment. These are linguistic questions. The growth in importance of analytics of all kinds in so many areas of business has increased the

> importance of domain-specific skills. But in between pure Mathematics and Statistics and domain-specific knowledge, there is a broad band of quite general knowledge that exploits the specific problems of meaning extraction and specific properties of text and language: In so many domains, unstructured text is where the information we need lies hidden. This is the area the proposed programs target.

(d) Describe community participation, if any, in the planning process. This may include prospective employers of graduates.

> This is process is ongoing. We have talked to senior technical staff at NTent, A-Life Medical, and Spawar, all of whom have employed students of ours at one time or another. We are hoping to combine our outreach efforts with those made by proposers of the new Big Data Master's, an effort being led by Ming-Hsiang Tsou of the Human Dynamics Center and Sam Shen of Computer Science (see the attached letter of support).

(e) Provide applicable workforce demand projections and other relevant data.

> The area of Data Science is a growing field. As Dr. Beck points out in his support letter, the U.S. Department of Labor Occupational Handbook predicts "the need for scholars and professionals [will] grow 15% through 2022". More immediately, a recent search of monster.com turned up over 1000 Data Science ads, a huge proportion of them in California, 272 of them in the San Diego area. kaggle.com, a more specialized site, turned up over 600 data science jobs, 60 of them specifically in the area of text analysis.

9.2 Student demand

(a) Provide compelling evidence of student interest in enrolling in the proposed program. Types of evidence vary and may include national, statewide, and professional employment forecasts and surveys; petitions; lists of related associate degree programs at feeder community colleges; reports from community college transfer centers; and enrollments from feeder baccalaureate programs, for example.

> Note that the programs being proposed do not entail a large investment of new resources. They mostly piggybacks off of existing courses that are succeeding. One new course is being proposed.
>
> Three of the core courses in the proposed certificate and minor are also courses in a basic certificate we now offer called the Computational Linguistics certificate: Ling 571, Ling 572, and Ling 581. Ling 571 and Ling 581 are long-standing courses which have had satisfactory enrollments for some time, in part because of the certificate. Ling 581 is cross-listed as a CS course (CS 581), and has historically

enrolled slightly better as CS 581 (primarily CS students) than as Ling 581 (primarily Ling students), in part because of its difficult material and programming requirement. It hit an enrollment high of 33 students in Spring 2015, with 18 of the students from CS. Many of the CS students who enroll are not certificate students but are simply looking for an elective.

The proposed certificate has one relatively new class that has been taught only 3 times, Ling 572. The first time Ling 572 was offered in Spring 2014, the class over-filled with 40 students, including students from Statistics, Geography, Computer Science, and Anthropology. Fall 2015 the course enrolled 24 students. The primary reasons for the drop in enrollments were (a) that an unusually large number of Stat students had enrolled in Spring 2014, due to a lower number of upper division and Grad Stat courses being offered that semester because of faculty leaves; and (b) Geography students have stopped enrolling because of a new Python course targeting GIS applications (taught as Geo 596). Interestingly, the remaining 24 students are quite diverse, 12 from Ling, 2 from Stat, 1 from CS, 1 from Econ student, 1 from MIS, and 1 from Spanish; there were also 6 Open University enrollments, indicating some interest in the San Diego community. Many of these students are being attracted by the existing Comp Ling certificate. We think if the new minor and New certificate are approved, we will continue to attract the diverse students coming now, but we will also attract a larger number of CES and Geography students pursuing this more technical certificate. Letters supporting this idea are attached, from Michael O'Sullivan (Chair) and Rich Levine in Math and Statistics, Leland Beck (Chair) in Computer Science, and Ming-Hsiang Tsou in Geography (Director of the Center for Human Dynamics in the Mobile Age). As Dr. Tsou points out in the attached letter of support, the content covered in these certificate courses also fits right in with the Big Data M.S. being proposed, so further down the road, we also expect students in that program to be either pursuing our certificate or taking certificate classes as electives.

The class whose near term success will be most closely tied to the success of the minor and proposed certificate is Ling 583, the new class being proposed with the new programs. This is envisioned as a capstone class that will take advantage of the deeper technical background of the students, and will be able to address how and why machine learning and statistical methods work, and what their limitations are. This class targets students who already have some expertise in data science, and are interested in acquiring specific expertise in the mining and analysis of text. The motivation for acquiring this specific expertise, then, is in part going to be tied to the market for jobs calling for it. We believe this market is growing

> rapidly, as discussed in the Societal and Public Needs Assessment section. As Dr. Beck points out in his support letter, the U.S. Department of Labor Occupational Handbook predicts “the need for scholars and professionals [will] grow 15% through 2022”. A recent search of monster.com turned up over 1000 Data Science ads, a huge proportion of them in California, 272 of them in the San Diego area. The kaggle.com site, with more specialized ads, turned up over 600 data science jobs, 60 of them specifically in the area of text analysis.

(b) Identify how issues of diversity and access to the university were considered when planning this program. Describe what steps the program will take to insure ALL prospective candidates have equitable access to the program. This description may include recruitment strategies and any other techniques to insure a diverse and qualified candidate pool.

> Linguistics is a discipline founded on the premise of diversity. It is critical that our inventory of analytical techniques be able to accommodate all languages, even and perhaps especially, endangered languages with dwindling populations. This is no less important in the age of information, when global expansion is bringing technology into areas with languages that may not have a long history of being written, and equally importantly, languages that may not have large bodies of annotated online text suitable for classical machine learning methods. The discipline needs both language experts with technical know-how and technical experts with linguistic sophistication. Accordingly, we plan (a) to add to our diverse body of existing linguistics students by trying to recruit students into the Minor from CAL language programs; (b) to focus our efforts on recruiting non-native speakers of English who are Statistics and CS majors, emphasizing that their existing language skills are marketable assets in this discipline.

(c) For master’s degree proposals, cite the number of declared undergraduate majors and the degree production over the preceding three years for the corresponding baccalaureate program, if there is one.

> NA.

(d) Describe professional uses of the proposed degree program.

> The proposed Minor/certificate will assist students in getting data science jobs in analytic consulting firms like Cyber Coders and NTent in San Diego, or analytic jobs in firms doing their own analysis in specific market areas, like the Nordstrom’s job described above or San Diego based A-Life Medical, which does analysis of the unstructured text found in medical records. There are also text analytics

jobs in the government sector, working for government organizations overwhelmed by text data, such as San Diego-based SPAWAR, or in intellgent analysis for organizations like the NSA.

(e) Specify the expected number of majors in the initial year; and three years and five years thereafter. Specify the expected number of graduates in the initial year, and three years and five years thereafter.

We expect 15 Minor/Certificate students the first year, with a growth rate of 20% for the next five years, culminating in 31 students 5 years out. Growth will be faster if the Big Data Master's is approved.

9.3 Existing support resources

A. Associated faculty

1) Jean Mark Gawron, Professor, Linguistics, Ph.D.
2) Robert Malouf, Associate Professor, Linguistics, Ph.D.
3) Richard Levine, Ph.D., Professor, Statistics, Ph.D.
4) Marie Roch, Professor, Computer Science, Ph.D.
5) Douglas Bigham, Assistant Professor, Linguistics, Ph.D.
6) Scott Kelley, Professor, Biology, Ph.D.

B. Facilities

The existing Computational Linguistics lab will be used for the courses in linguistics. In addition to basic texts, many course materials will be in interactive online environments, such as Python notebooks.

C. Accessibility

Course materials will be made available on the web. The Comp Ling Lab is in Storm Hall West, which is wheel chair accessible.

D. Technology

There are 15 PCs running Linux with Python and R (the primary programming languages of the minor) installed in the computational linguistics lab. In addition, both R and Python are freely distributed software and students can acquire local copies on their home machines.

10 Plans

We will be proposing a Topics version of Ling 583 in February 2016. and it will first be taught in the Fall of 2016.

We will propose a graduate course and an advanced certificate if enrollments justify the additional resources.