Text Modeling for the Humanities and Social Sciences

CS7180: Special Topics in AI

David A. Smith

Northeastern University
NULab for Texts, Maps, and Networks
Digital Breadcrumbs

- Email
- Text messaging
- Social media
- Phone records
- Web links
- Web searches
- Smart cards
- License plates

*Slide from David Lazer*
Digital Detritus

- Tweets
- Emails of government employees
- Supreme Court decisions
- Anti-vax message boards
- Newspaper ads for runaway slaves
- Nineteenth-century novels
- Letters written by seventeenth-century scientists
The Roots of Big Data

- Big Government, Business, Science
- Social change: Living online
- Digitizing the past
Socrates: I heard, then, that at Naucratis, in Egypt, was one of the ancient gods of that country ... and the name of the god himself was Theuth. He it was who invented numbers and arithmetic and geometry and astronomy, also draughts and dice, and, most important of all, letters. Now the king of all Egypt at that time was the god Thamus ... To him came Theuth to show his inventions, saying that they ought to be imparted to the other Egyptians. But Thamus asked what use there was in each, and as Theuth enumerated their uses, expressed praise or blame, according as he approved or disapproved. The story goes that Thamus said many things to Theuth in praise or blame of the various arts, which it would take too long to repeat; but when they came to the letters, “This invention, O king,” said Theuth, “will make the Egyptians wiser and will improve their memories; for it is an elixir of memory and wisdom that I have discovered.” But Thamus replied, “Most ingenious Theuth, one man has the ability to beget arts, but the ability to judge of their usefulness or harmfultness to their users belongs to another; and now you, who are the father of letters, have been led by your affection to ascribe to them a power the opposite of that which they really possess. For this invention will produce forgetfulness in the minds of those who learn to use it, because they will not practice their memory. Their trust in writing, produced by external characters which are no part of themselves, will discourage the use of their own memory within them. You have invented an elixir not of memory, but of reminding; and you offer your pupils the appearance of wisdom, not true wisdom, for they will read many things without instruction and will therefore seem to know many things, when they are for the most part ignorant and hard to get along with, since they are not wise, but only appear wise.

Phaedrus: Socrates, you easily make up stories of Egypt or any country you please.
Orality and Literacy

Walter J. Ong
The printing press as an agent of change
COMPLETE IN ONE VOLUME
Elizabeth L. Eisenstein
SEEING LIKE A STATE

How Certain Schemes to Improve the Human Condition Have Failed
IMAGINED COMMUNITIES

BENEDICT ANDERSON

'A brilliant exegesis on nationalism'
The Nation

NEW EDITION
nothing but the


events, unless the accounts from many quarters as to General
Schenck's instructions are utterly belied, the new American
Ambassador will bring us quite reasonable, though not
perhaps wholly admissible demands,—demands which we
certainly ought to consider most gravely, and of which we
should do well to yield frankly and freely all that we should
ourselves feel called upon, in the same circumstances, to
press. If we do so, General Schenck's mission may make
England safer and stronger than she has ever been since the
close of the Civil War in 1865, and will give her a reputation
for moderation and candour as well.

ENGLISH PUBLIC OPINION ON THE WAR.

Some of the philosophers should turn their attention from
the subject of spectroscopic investigations and the invention
of electrometers, galvanometers, hygrometers, and so forth,
to the far more difficult problem of inventing a mode of
measuring the intensity and diffusion of political wishes and
convictions. No task at present is more difficult for a States-
man than this. There are, indeed, all sorts of shades of dif-
ference between the character of really prevalent and prepo-
derant public opinions, of which no man, however acute,
ever forms more than a purely conjectural impression, and of
which, nevertheless, any respectfully-accurate measure would
be a matter of the highest political importance. For in-
stance, there is at times a public opinion on one side of a
question which is very widely diffused, but of very slight in-
tensity,—which, in fact, amounts to nothing more than
a wish in a particular direction without a will,
and still more without any intention of submitting
to a considerable sacrifice rather than not carry
out the will into action. Again, there is such a thing as
But such work is, of course, only a beginning in exploring a territory that to all intents and purposes still lies largely unexplored. So, broadly, the challenge offered by this book to historians and social scientists is one that still has to be taken up. If they do so, they will find it is not enough to rely on the old traditional methods of either history or the social sciences — and least of all on the arid abstractions of the ‘structuralists’ or the antiquated crowd-psychology of Le Bon and his followers, which dismiss rioters without discrimination as ‘mobs’ and see ‘conspiracy’ round every corner. Statistical methods certainly have their place; and the much-abused computer may, in such cases, serve as an excellent tool for classifying, correlating, enumerating and bringing order to material which, if voluminous enough, may easily get out of hand. Moreover, the techniques of the geographer may be indispensable for plotting the topography of a complex movement and might, even in the present instance, have helped to answer more convincingly such questions as to why the Great Fear struck some districts and not others. So the historian of similar movements requires such help as he can get from his colleagues in the other social sciences. But, equally, there can be no substitute for the intense industry and patient zeal of the professional historian, which enables him (as Lefebvre does here) to chart in precise detail the course of a movement by noting the days, and even the hours, of its appearance: by such means alone can the old bogey of rampaging ‘mobs’ and ubiquitous ‘conspiracies’ be laid to rest, or at least be reduced to proper historical proportions. Let
Build a DH Project with this One Weird Trick
The Return of the Editor
The Rise of Big Data
Where is the Editor?
Case Study
Books as Big Data
The Internet Archive and Open Library offers over 10,000,000 fully accessible and texts. There is also a collection of 300,000 modern eBooks that may be borrowed or downloaded by the print-disabled at OpenLibrary.org.
Bernard Mandeville  The Fable of the Bees  1729

Law, William, 1686-1761  Remarks on The fable of the bees  1844
Maurice, Frederick  Denison, 1805-1872

Irwin Primer  The Fable Of The Bees Or Private Vices Publick Benefits  1962

Bernard Mandeville  The Fable of the Bees : Or, Private Vices, Publick Benefits: With an Essay on Charity and ...  1806
<table>
<thead>
<tr>
<th>Authors</th>
<th>Title</th>
<th>Year</th>
</tr>
</thead>
<tbody>
<tr>
<td>Shelley, Mary</td>
<td>Frankenstein</td>
<td>1993</td>
</tr>
<tr>
<td>Wollstonecraft, 1797-1851</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mary Wollstonecraft, 1797-1851</td>
<td>Frankenstein: or, The modern Prometheus</td>
<td>—</td>
</tr>
<tr>
<td>Shelley</td>
<td>Frankenstein: or, The modern Prometheus</td>
<td>1869</td>
</tr>
<tr>
<td>Wordsworth Collection</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mary Shelley</td>
<td>Frankenstein, Or, The Modern Prometheus</td>
<td>1823</td>
</tr>
<tr>
<td>Kenneth Branagh</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mary Wollstonecraft, 1797-1851</td>
<td>Frankenstein: Or, The Modern Prometheus</td>
<td>1869</td>
</tr>
<tr>
<td>Shelley</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mary Wollstonecraft, 1797-1851</td>
<td>Frankenstein: Or, The Modern Prometheus</td>
<td>1891</td>
</tr>
</tbody>
</table>
The catalogue says this is Latin.
Conservatively, 1% = 32,000 archive.org books tagged w/wrong language
Metadata and Structure inside the Book
I have preferred a reference in a footnote, and have not included the passages in my final summary.

The two word lists are important as showing, as Gildersleeve\(^1\) points out, that Persius "has enriched his vocabulary from Lucilius's store of drastic [I should perhaps be inclined to substitute popular] words," and that even where there is no question of conscious imitation, the mind of Persius was so suffused with the verses, the diction, the arguments of Lucilius, that they became a part of the texture of his satire.

Finally, I desire to call attention to the cumulative force of an argument, based on the precise testimonia of the ancients, and on the comparative study of the imitations, and vocabularies of Lucilius and Persius. In view of the fragmentary condition of Lucilius, I feel that the mass of evidence becomes all the more striking, and forces us to the conclusion evidently made by the ancients themselves, that Lucilius is a source for Persius second only to Horace in importance.

ADDENDUM

Just as this article goes to press the "Persius Probleme," \textit{Wiener Studien}, xxxi, 125-135, 233-243, and "Persius und Lucilius," B. 244-249, of Emil Gaar become accessible to me. I entirely agree to Gaar's argument for the close connection between the prologue and the first satire. Cf. above, pp. 141-142. To my mind the relation between the 8 choliambics and the 14 hexameters in Petronius, § 4, affords material confirmation for this same argument.

I find myself unable to subscribe to Gaar's view of the relation between the tenth book of Lucilius and the prologue. In the first place, Buecheiner's article, \textit{Rh. Mus.} xxxiv, 287, proves rather a connection between the first satire and the tenth book. Moreover, the scanty surviving fragments of Book x in no way suggest the prologue. In the second place, though unquestionably the scholiast on line 2, and the Persius \textit{vita}, as Gaar convincingly shows, refer to different passages, all the facts of the case are satisfied by the inference (1) that the scholiast was referring to Persius, l. 1; (2) that \textit{sit} \textit{primo} \textit{maxim} \textit{omnibus} \textit{detractusus} of the \textit{vita} refers to the general tone of the first satire. Thus the \textit{detectatio sui} would be lines 1-4, followed by the \textit{inspectatio postermal et oratorium}. At least the scholiast felt this, for he says, l. 1: \textit{Semetipsum} redarguit, quod ipse reliquit carmina, quae volgis lectorum non sit, quoniam non sint vulgaria, etc.

\(^{1}\) Introduction to Persius, pp. xxiii and xxv.
I have preferred a partial reference, rather than to have included the passage in its entirety. The two word limits on the caption sleeve points out, the editor of Lucilius's works (G. B. de Piles) says "to substitute popular words," and that even where there is no question of conscious imitation, the mind of Persius was so suffused with the verses, the diction, the arguments of Lucilius, that they became a part of the texture of his satire.

Finally, I desire to call attention to the cumulative force of an argument, based on the precise testimonia of the ancients, and on the comparative study of the imitations, and vocabularies of Lucilius and Persius. In view of the fragmentary condition of Lucilius, I feel that the mass of evidence becomes all the more striking, and forces us to the conclusion evidently made by the ancients themselves, that Lucilius is a source for Persius second only to Horace in importance.

**ADDENDUM**

Just as this article goes to press the “Persius Probleme,” Wiener Studien, xxxi, 128-135, 233-243, and “Persius und Lucilius,” B. 244-249, of Emil Gaar become accessible to me. I entirely agree to Gaar’s argument for the close connection between the prologue and the first satire. Cf. above, pp. 141-142. To my mind the relation between the 8 choliambics and the 14 hexameters in Petronius, § 4, affords material confirmation for this same argument.

I find myself unable to subscribe to Gaar’s view of the relation between the tenth book of Lucilius and the prologue. In the first place, Buecheler’s article, Rh. Mus., xxxix, 287, proves rather a connection between the first satire and the tenth book. Moreover, the scanty surviving fragments of Book x in no way suggest the prologue. In the second place, though unquestionably the scholiast on line 2, and the Persius vita, as Gaar convincingly shows, refer to different passages, all the facts of the case are satisfied by the inference (1) that the scholiast was referring to Persius, l. 1; (2) that the sibi primo max omnibus detracturum of the vita refers to the general tone of the first satire. Thus the detractatio sibi would be lines 1-4, followed by the insectatio posturum et oratorum. At least the scholiast felt this, for he says, l. 1: Sometiorem redarguit, quid ipse relinquuit carmina, quae volgus lecturam non sit, quoniam non sint vulgaria, etc.

1 Introduction to Persius, pp. xxiii and xxv.
New article and author

Quotation (within a quotation), in Latin, from earlier than 1909
By being in his eye. Most gracious duke,  
To my unfolding lend your prosperous ear;  
And let me find a charter in your voice,  
To assist my simpleness.

_Duke._ What would you, Desdemona?

_Desdemona._ That I did love the Moor to live with him,  
My downright violence and storm of fortunes  
May trumpet to the world: my heart's subdued  
Even to the very quality of my lord:

I saw Othello's visage in his mind,  
And to his honours and his valiant parts  
Did I my soul and fortunes consecrate.  
So that, dear lords, if I be left behind,  
A moth of peace, and he go to the war,  
The rites for which I love him are bereft me,  
And I a heavy interim shall support  
By his dear absence. Let me go with him.

_Othello._ Let her have your voices.

Vouch with me, heaven, I therefore beg it not,  
To please the palate of my appetite,  
Nor to comply with heat—the young affects  
In me defunct—and proper satisfaction,  
But to be free and bounteous to her mind:  
And heaven defend your good souls, that you think  
I will your serious and great business scant  
For she is with me. No, when light-wing'd toys  
Of feather'd Cupid seel with wanton dulness  
My speculative and offe'd instruments,  
That my disports corrupt and taint my business,  
Let housewives make a skilet of my helm,  
And all indign and base adversities  
Make head against my estimation!

_Duke._ Be it as you shall privately determine,  
Either for her stay or going. The affair cries haste,  
And speed must answer it.
Text & Commentary

Aeneid books 1-3

Commentary on Aeneid 1-3, with matching lemmata in green
Aligning Reference Texts
Cf. work with Ryan Cordell & Elizabeth Dillon on Viral Texts
C19 newspapers were an all-purpose medium
C19 newspapers were built on systems of textual sharing.
THE RESTING PLACE.

BY CHARLES MACKAY.

Tell me, ye winged winds,
That round my pathway roar,
Do ye not know some spot
Where mortals weep no more?
Some lone and pleasant dell,
Some valley in the west,
Where free from toil and pain,
The weary soul may rest?

Tell me ye winged winds,
That round my pathway roar,
Do ye not know some spot
Where mortals weep no more?
Some lone and pleasant dell,
Some valley in the west,
Where free from toil and pain,
The weary soul may rest?

Tell me ye winged winds,
That round my pathway roar,
Do ye not know some spot
Where mortals weep no more?
Some lone and pleasant dell,
Some valley in the west,
Where free from toil and pain,
The weary soul may rest?

The loud wind dwindled to a whisper low,
And sighed for pity as it answered "No!"
"Mother wants to know if you won't lend her your preserving-kettle, 'cause she wants to use up some apple-sauce." "I would with pleasure, my boy, but the truth is, the last time we lent it to your mother she preserved it so effectually that we haven't seen it since." "Well, you needn't be so sassy about your old kettle. Guess it was full of holes when we borrowed it, and mother wouldn't have troubled you again, only we see'd you bring home a new one."

Jinks
International Atomic Energy Agency has repeatedly called attention to Iran's unlawful nuclear activities, and, as a result, the United Nations Security Council has adopted a range of sanctions designed to encourage the Government of Iran to cease those activities and comply with its obligations under the Treaty on the Non-Proliferation of Nuclear Weapons (commonly known as the "Nuclear Non-Proliferation Treaty").

The International Atomic Energy Agency (IAEA) has repeatedly called attention to the illicit nuclear activities of the Islamic Republic of Iran, and, as a result, the United Nations Security Council has adopted a range of sanctions designed to encourage the Government of the Islamic Republic of Iran to cease those activities and comply with its obligations under the Treaty on Non-Proliferation of Nuclear Weapons, done at Washington, London, and Moscow July 1, 1968, and entered into force March 5, 1970 (commonly known as the "Nuclear Non-Proliferation Treaty").
The language of some household song,
The perfume of some cherished flower,
Though gone from outward sense, belong
To memory's after hour.
So with our words, or harsh, or kind,
Uttered, they are not all forgot,
They leave their influence on the mind,
Post on, but perish not;
As they are spoken, so they fall
Upon the spirit spoken to,
Scorch it like drops of burning gall,
Or soothe like honey-dew.
So with our deeds, for good or ill,
They have their power, scarce understood;
Then let us use our better will
To make them ripe with good.
Like
They leave their influence on the mind,
Evening star. (Washington, D.C.) 1854-1972
So with our deeds, for good or ill,
Uttered, they are not all forgot.

Or sooth like honey-dew.
Scorch it like drops of burning gall,
Then let us see our better will
They have their power scarce understood;
So with our deeds, for good or ill,
Uttered—they are not all forgot;
To memory’s after hour.

To make them rife with good:
Like
Class Logistics

• In each class, we will read and discuss ~2 papers.

• **One** student will be responsible for leading discussion of each paper,
  • including background on methods used, and
  • suggestions for extending the work.

• Before class, **all** students should post on Piazza short paper summaries and discussion questions.
Projects

• Students will complete course projects and papers
• Papers intended to be a (part of a) new publication
• Possible coauthorship with clearly assigned credit
• Topics include: new models, applying models to new datasets, replicating state-of-the-art models
Projects

- Discuss initial project ideas on October 5
- Present project outline on November 9
- Final papers due December 5
- Final presentations on December 7
- Etymology is not destiny (sorry, early Plato), but a seminar is a seedbed
Grades

- Final project paper: 80%
- Participation and presentations: 20%
- Collaborative projects must specify a collaboration plan and assignment of credit
Discussion

- Research background?
- Humanities and social sciences experience?
- Interesting problems?
- Interesting data?
Course Topics

- Representing text: Tokenization
- Word vectors and distributed representations
- Text categorization
- Language models and topic models
- Dynamic models and temporal change
- Entity and relation extraction
Course Topics

• Plot and character
• Dialogue and argumentation
• Language and power relations
• Geographical and social variation
• Genre
• Information cascades
Course Topics

- Document analysis and recognition
- Speech analysis and recognition
- Community structure and communication
- Framing and rhetoric
- Laws and treaties
- Criminology
Course Topics

• Censorship
• Economic history
• Stylometry
• Citation analysis
• Collaborative editing
Next Class

