LING 120, Fall 2017: Language and Computers

Topic: Overview of Natural Language Processing

Instructor: Sowmya Vajjala

Iowa State University, USA

13 October 2017 (Week 8)

Class outline

- Midterms: comments
- Assignment 3 discussion
- General Review and discussion of topics
- Feedback on the course so far.

Midterms: Comments

- 6 teams: voice recognition (1), duo lingo (3), regular expressions (2)
- Good show - perfect scores! (to be fair, I also did not nitpick!)
- Content: Lot of new stuff beyond what was discussed in the class. Hopefully, all of you enjoyed as much as I did!
- presentation issues: maintaining co-ordination between presenters, eye-contact with audience, taking care about what is put on slides, sticking to time etc. (those can improve with practice and experience!)

Voice Recognition

- Loved those spectrogram images!
- Good observations about Dictation tool.
- Wishlist: someone doing this for multiple languages and comparing accuracy of speech recognition for some non-English language (with accents!)

Duolingo presentations

- Duo Lingo - different languages, exercise sequences, advanced level exercises, writing systems, comparison with classroom learning, grammar
- languages discussed: English, German, Greek, Chinese, Hebrew, Italian, Japanese, Russian, Spanish, Ukranian
- Hopefully, you learnt more about tutoring systems than what is in the textbook!

Regular Expressions Presentations

- Searching through code, searching in MS Word with Macros.
- Nice examples/videos/demos.
- A little bit of complex stuff for non-CS people, something you could work on in terms communicating to a broader audience.

Assignment 3 discussion

- 1 (a): Duolingo vs TAGARELA
- 1 (b): British council test - not adaptive (most of you - perhaps the ones who tried it twice) got it right.
- 2 (a): "he", "assembly", 3170. Most diverse answers are here :-)
- 2 (b): Amazon vs the search engine X.

General Remarks

- When you don't understand something, ask.
- Office hours: Monday, Wednesday 1–2 pm. Ross 331

General Remarks

- When you don't understand something, ask.
- Office hours: Monday, Wednesday 1–2 pm. Ross 331
- If you don't tell me, I cannot know you are having difficulties.
- Participate in the class, or online, or talk to me in person if you are too shy.

General Remarks

- When you don't understand something, ask.
- Office hours: Monday, Wednesday 1–2 pm. Ross 331
- If you don't tell me, I cannot know you are having difficulties.
- Participate in the class, or online, or talk to me in person if you are too shy.
- It is just another course, yes, but why lose grade in a 100-level course unnecessarily?

General Review of Topics so far

We finished half of the book (and more!!!)

1. Encoding and inputting text and speech on computers
2. Writers aids - the only topic which no one chose for midterms :-)
3. Language tutoring systems
4. Searching
5. Overview of natural language processing

(3 assignments and 1 mid-term done!)

Whats in store next?

1. Quick overview of language and cryptography
2. automatically classifying text (e.g., spam email classification)
3. Dialog systems, speech based human-machine interaction
4. Machine translation

Next week

- Topic: (mainly) Text classification
- Readings: Chapter 5 in the textbook
- Deadlines: Assignment 4 (21st October)

What is text classification?

A randomly picked student should answer each question!

- if I say spam email classification as an example, what are some other such practical applications you can think of?

What is text classification?

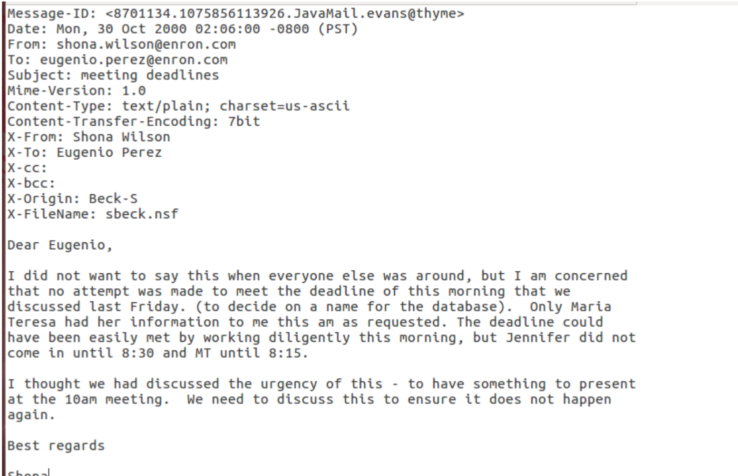
A randomly picked student should answer each question!

- if I say spam email classification as an example, what are some other such practical applications you can think of?
- How easy is it to do spam classification for humans?

What is text classification?

A randomly picked student should answer each question!

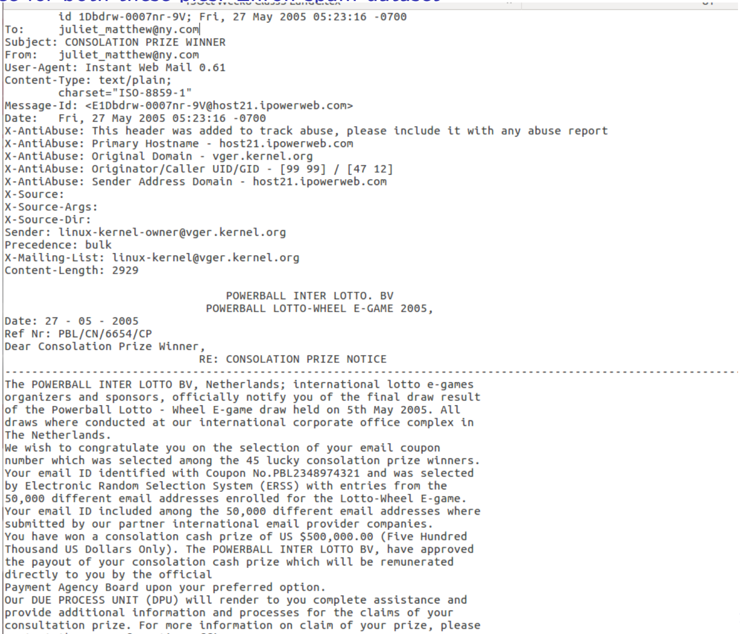
- if I say spam email classification as an example, what are some other such practical applications you can think of?
- How easy is it to do spam classification for humans?
  Is this spam?



```
Message-ID: <8701134.1075856113926.JavaMail.evans@thyme>
Date: Mon, 30 Oct 2000 02:06:00 -0800 (PST)
From: shona.wilson@enron.com
To: eugenio.perez@enron.com
Subject: meeting deadlines
Mime-Version: 1.0
Content-Type: text/plain; charset=us-ascii
Content-Transfer-Encoding: 7bit
X-From: Shona Wilson
X-To: Eugenio Perez
X-cc:
X-bcc:
X-Origin: Beck-S
X-FileName: sbeck.nsf

Dear Eugenio,

I did not want to say this when everyone else was around, but I am concerned
that no attempt was made to meet the deadline of this morning that we
discussed last Friday. (to decide on a name for the database).  Only Maria
Teresa had her information to me this am as requested. The deadline could
have been easily met by working diligently this morning, but Jennifer did not
come in until 8:30 and MT until 8:15.

I thought we had discussed the urgency of this - to have something to present
at the 10am meeting.  We need to discuss this to ensure it does not happen
again.

Best regards

Shona
```

Is this spam?

Source for both these pics: Enron spam dataset



```
        id 1Dbdrw-0007nr-9V; Fri, 27 May 2005 05:23:16 -0700
To:     juliet_matthew@ny.com
Subject: CONSOLATION PRIZE WINNER
From:   juliet_matthew@ny.com
User-Agent: Instant Web Mail 0.61
Content-Type: text/plain;
        charset="ISO-8859-1"
Message-Id: <E1Dbdrw-0007nr-9V@host21.ipowerweb.com>
Date:   Fri, 27 May 2005 05:23:16 -0700
X-AntiAbuse: This header was added to track abuse, please include it with any abuse report
X-AntiAbuse: Primary Hostname - host21.ipowerweb.com
X-AntiAbuse: Original Domain - vger.kernel.org
X-AntiAbuse: Originator/Caller UID/GID - [99 99] / [47 12]
X-AntiAbuse: Sender Address Domain - host21.ipowerweb.com
X-Source:
X-Source-Args:
X-Source-Dir:
Sender: linux-kernel-owner@vger.kernel.org
Precedence: bulk
X-Mailing-List: linux-kernel@vger.kernel.org
Content-Length: 2929

                              POWERBALL INTER LOTTO. BV
                           POWERBALL LOTTO-WHEEL E-GAME 2005,
Date: 27 - 05 - 2005
Ref Nr: PBL/CN/6654/CP
Dear Consolation Prize Winner,
                             RE: CONSOLATION PRIZE NOTICE
------------------------------------------------------------------------------------------------------------
The POWERBALL INTER LOTTO BV, Netherlands; international lotto e-games
organizers and sponsors, officially notify you of the final draw result
of the Powerball Lotto - Wheel E-game draw held on 5th May 2005. All
draws where conducted at our international corporate office complex in
The Netherlands.
We wish to congratulate you on the selection of your email coupon
number which was selected among the 45 lucky consolation prize winners.
Your email ID identified with Coupon No.PBL2348974321 and was selected
by Electronic Random Selection System (ERSS) with entries from the
50,000 different email addresses enrolled for the Lotto-Wheel E-game.
Your email ID included among the 50,000 different email addresses where
submitted by our partner international email provider companies.
You have won a consolation cash prize of US $500,000.00 (Five Hundred
Thousand US Dollars Only). The POWERBALL INTER LOTTO BV, have approved
the payout of your consolation cash prize which will be remunerated
directly to you by the official
Payment Agency Board upon your preferred option.
Our DUE PROCESS UNIT (DPU) will render to you complete assistance and
provide additional information and processes for the claims of your
consultation prize. For more information on claim of your prize, please
```

Spam in SMS: What is spam?

- "Did you catch the bus ? Are you frying an egg ? Did you make a tea? Are you eating your mom's left over dinner ? Do you feel my Love ?"
- "Forwarded from 44871240400: Please CALL 08712404000 immediately as there is an urgent message waiting for you."
- "Do you realize that in about 40 years, we'll have thousands of old ladies running around with tattoos?"
- "Realy sorry-i don't recognise this number and am now confused :) who r u please?! "

source: online spam dataset. `https://goo.gl/7Dq85S`

Spam vs Ham - is it so easy?

- Not always. A job ad is perhaps not spam. But if the same ad comes repeatedly, it is.
- A group email is not a spam for someone. But it is, for others.
- Is it easier or difficult with SMS compared to email?

Spam vs Ham - is it so easy?

- Not always. A job ad is perhaps not spam. But if the same ad comes repeatedly, it is.
- A group email is not a spam for someone. But it is, for others.
- Is it easier or difficult with SMS compared to email?
- Are you satisfied with your email provider's spam classification?

Spam vs Ham - is it so easy?

- Not always. A job ad is perhaps not spam. But if the same ad comes repeatedly, it is.
- A group email is not a spam for someone. But it is, for others.
- Is it easier or difficult with SMS compared to email?
- Are you satisfied with your email provider's spam classification?
- What do you think it is doing?

Feedback about the course

- Please fill up the form given (all the 4 blocks)
- whats in it for you: you can see some improvements for the rest of the semester
- whats in it for me: feedback to improve teaching practices
- whats in it for future students: a better designed course