Carnegie Mellon University
Language
Technologies
Institute

11-411/11-611 Natural Language Processing

Introduction

David R. Mortensen and Lori Levin

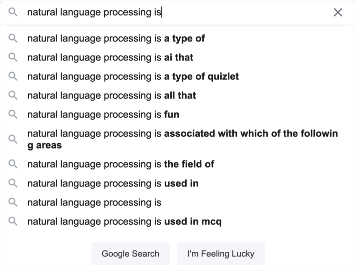
January 17, 2023

Language Technologies Institute

What is NLP?

NLP is Everywhere

Did you ever wonder how web search engines work...



natural language processing is

- natural language processing is **a type of**
- natural language processing is **ai that**
- natural language processing is **a type of quizlet**
- natural language processing is **all that**
- natural language processing is **fun**
- natural language processing is **associated with which of the following areas**
- natural language processing is **the field of**
- natural language processing is **used in**
- natural language processing is
- natural language processing is **used in mcq**

...or how Google can anticipate what you're searching for?

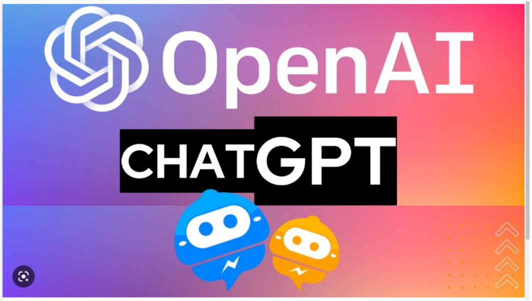
That's NLP!

NLP is Everywhere

Did you ever wonder how digital assistants work?

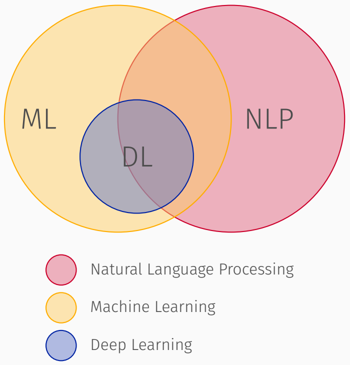
That's NLP!

ChatGPT is also NLP!

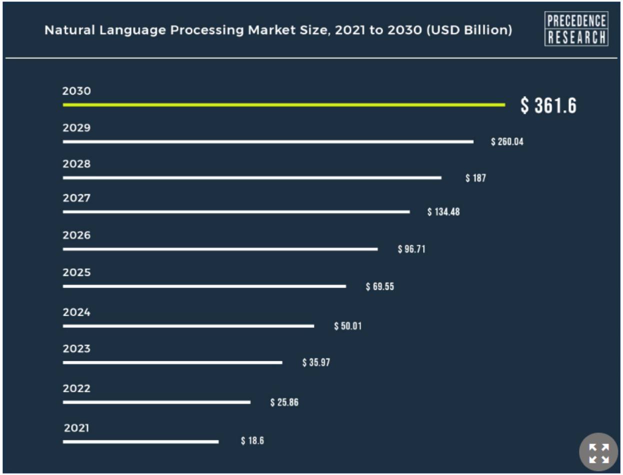


NLP is the Computational Analysis and Synthesis of Speech and Language

- NLP is one of the most important AI fields today
- Engineering focus—solving practical problems
- Scientific focus
    - the last bit that eludes the state of the art
    - new frontiers like common sense reasoning

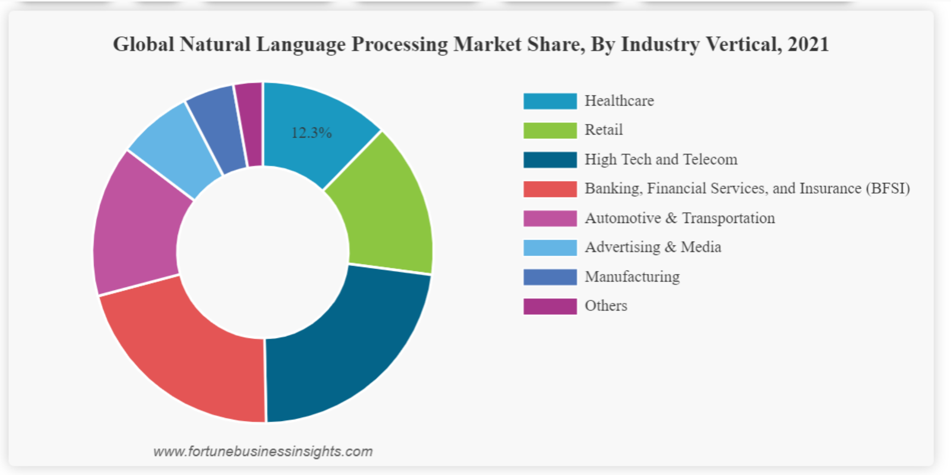


Societal Impact: NLP is a big market affecting many aspects of daily life



https://www.globenewswire.com/en/news-release/2022/09/14/2516050/0/en/Natural-Language-Processing-Market-Size-to-Expand-USD-361-6-Bn-by-2030.html

Societal Impact: NLP is a big market affecting many aspects of daily life



https://www.fortunebusinessinsights.com/industry-reports/natural-language-processing-nlp-market-101933

Societal Impact

- Virtual Agents and Chatbots
- Text and speech analytics
  - Turning unstructured data into insights, trends, and patterns
  - Fraud detection
- Communication with Robots
  - Military, medical, dangerous places, etc.
- Optical Character Recognition
- Customer satisfaction (sentiment detection)

https://www.fortunebusinessinsights.com/industry-reports/natural-language-processing-nlp-market-101933
https://www.fortunebusinessinsights.com/chatbot-market-104673

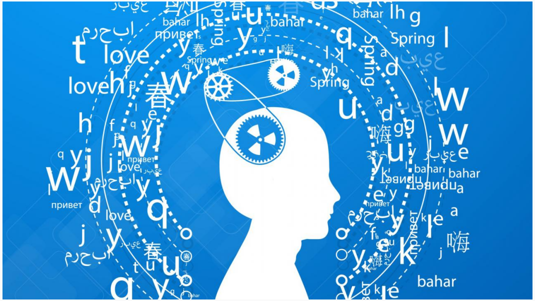
https://www.fortunebusinessinsights.com/text-analytics-market-104813

We Will Approach NLP at Multiple Levels in Each Lecture

Applications	The user-facing end goals of NLP, such as machine translation, search, question answering, "autocomplete", etc.,
Tasks	The abstract, computationally well-defined building blocks of applications (like sequence transduction, classification, sequence labeling, etc.,)
Methods	The concrete algorithms and technical approaches that are used to perform tasks and therefore construct applications (like using transformers to perform a language modeling task in a question generation system)

NLP Is Different from Computational Linguistics

- Computational linguistics is the scientific study of language using computers
- Closely related to NLP, but with different aims
- Many techniques and datasets used in NLP come from computational linguistics
- Computational linguistics uses many techniques from NLP



You Will Learn Some Linguistics in this Course

- NATURAL LANGUAGE PROCESSING is processing **natural language**
- The science that studies **natural language** is LINGUISTICS
- If you were engaged in computational modeling of physical systems, you would benefit from knowing some physics
    - You would know the difference between folk-science and real science.
    - In folk-physics, people think things dropped from moving airplanes fall straight down. Would you settle for that?
- Knowing basic linguistics will make you a deeper NLP thinker
    - Some of what you think you know about natural language is folk-science.
    - Compared to other sciences, it is relatively recent that people figured out how to study natural language as a science (as opposed to an art or a dogmatic grammar lesson). It hasn't made it into schools yet. But you will start to learn it here.

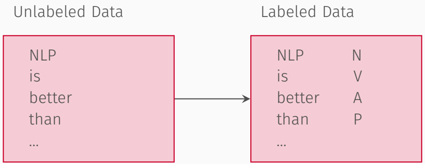
Knowing deep learning tells you what NLP can do now

Knowing linguistics tells you what NLP could do, but doesn’t yet

A Technical Framework for This Course

Supervised Machine Learning

TASK



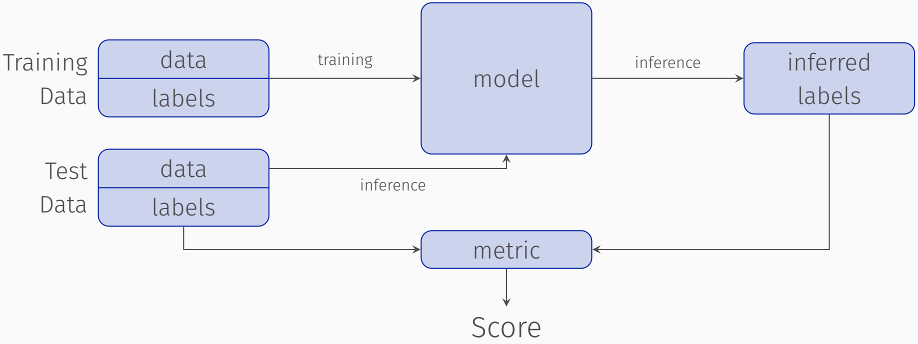
Unlabeled Data

NLP
is
better
than
…

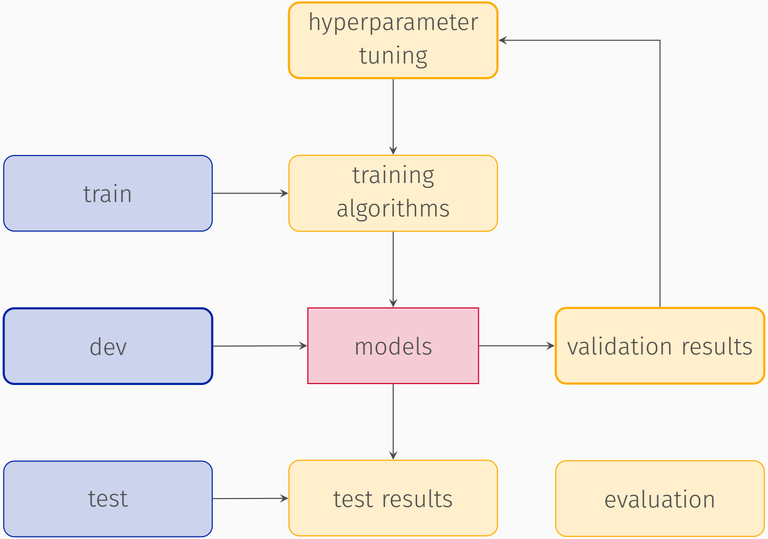
Labeled Data

NLP	N
is	V
better	A
than	P
…	

METHOD



The Validation or Development Set



Course Objectives and Overview

Learning Objectives

At the end of this course, a student will be able to:

- Implement a range of fundamental NLP algorithms for question answering, classification, language modeling, sequence labeling, and sequence transduction.
- Leverage morphology, syntax, semantics, discourse, and speech to perform NLP tasks.
- Recognize the class of tasks to which a specific natural language task belongs and identify an appropriate type of model to solve it.
- Design and develop a medium-scale NLP project with multiple components.

Module 1: Question Answering

Context: Another approach to brain function is to examine the consequences of damage to specific brain areas. Even though it is protected by the skull and meninges, surrounded by cerebrospinal fluid, and isolated from the bloodstream by the blood–brain barrier, the delicate nature of the brain makes it vulnerable to numerous diseases and several types of damage.

Question: What does the skull protect?

Answer: ”text”: [”brain”], ”answer-start”: [280]

Question: What sare the benifts of the blood brain barrir?

Answer: text”: [”isolated from the bloodstream”], ”answer-start”: [195]

Module 1: Question Answering

At the end of this module, students will be able to:

- Build a basic question answering system
- Annotate data for use in an NLP task
- Start on the semester-long project on QA

Module 2: Computational Morphology

Example:

Suppose you are building a search engine. When people look up a word like “computer”, say because they want to buy one, they might want to retrieve the following:

Dell **computer**, $500
We sell **computers** cheap
The best in **computing** machinery at great prices

Module 2: Computational Morphology

At the end of this module, students will be able to:

- Identify how the word-structure of a language will interact with NLP tasks by employing basic morphological concepts (root, prefix, suffix, lexeme, inflection, derivation)
- Build a lemmatizer for English verbs

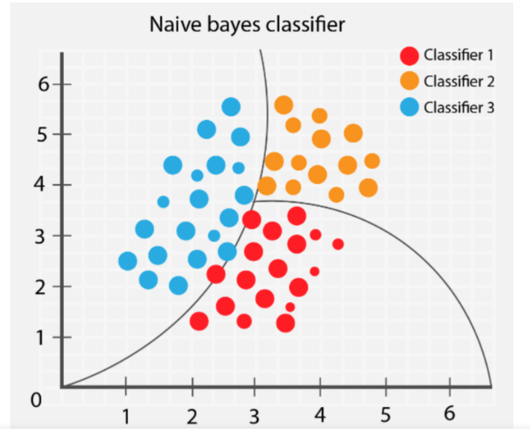
Module 3: Classification

Consider the following tweets about Zolgar corporation:

- "Zolgar corporation makes the best widgets I have ever used."
- "I'll never buy another Zolgar doodad."
- "Zolgar thingamagigs are not the worst I've tried."

Which of these tweets are positive and which are negative?

Module 3: Classification



At the end of this module, students will be able to:

- Characterize kinds of classification tasks in NLP
- Implement naive Bayes and logistic regression classifiers
- Build a basic language identification model

Module 4: Language Models

What is a language model?

A language model computes the probability of a next word given the history of previous words.

Example:
“I attend Carnegie Mellon.” is more probable than
“I attend Carnegie hippopotamus”.
Famous language models that you can use off the shelf:

- BERT
- GPT3
- T5

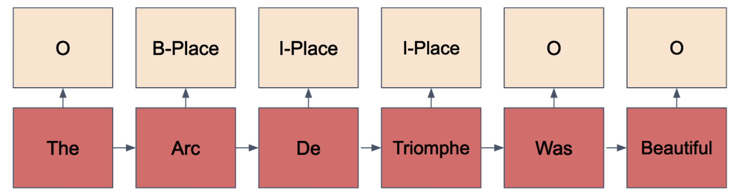
Module 4: Language Models

At the end of this module, students will be able to:

- Articulate what a language model does and the role it plays in NLP tasks
- Implement an n-gram language model
- Describe how neural language models work and what their relationship is to word (and sentence) embeddings

Module 5: Sequence Labeling Models

You sometimes want to find all of the names in a collection of documents. You can label each word based on whether it is at the beginning of a name, inside a name, or outside a name.



This is called NAMED ENTITY RECOGNITION and is a kind of SEQUENCE LABELING.

Module 5: Sequence Labeling Models

At the end of this module, students will be able to:

- Determine which NLP tasks can be reduced to sequence labeling
- Implement an HMM, the simplest practical sequence labeling model
- Use conditional random fields (CRFs) for sequence labeling
- Develop an HMM-based sequence-labeling model for part of speech (POS) tagging

Module 6: Deep Neural Networks for Sequence Processing

Suppose you want to take the paragraph:

> *My favorite course during Fall 2022 was Natural Language Processing, offered at Carnegie Mellon University. It was fun, exciting, and opened the doors of an exciting field to me. I learned to develop complex NLP systems and to use pretrained models to achieve a variety of tasks.*

And get back the summary:

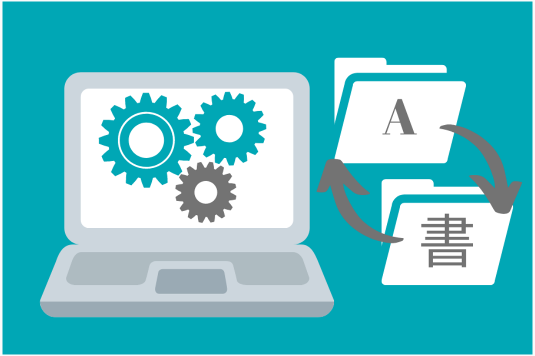
> *Natural Language Processing was a good course and I learned a lot in it.*

In this case you are converting (transducing) one sequence of words into another.

Module 6: Deep Neural Networks for Sequence Processing

At the end of this module, students will be able to:

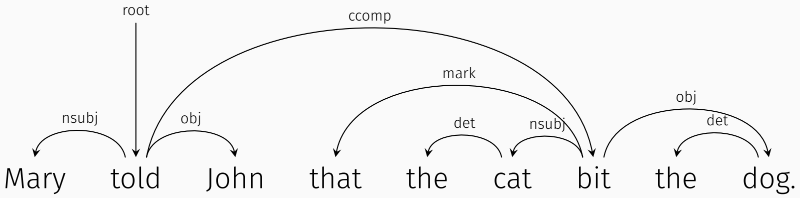
- Implement a sequence-to-sequence model using existing toolkits
- Train and test a simple machine translation model using PyTorch



Module 7: Syntax

At the end of this module, students will be able to:

- Identify two approaches to computational syntax: constituency and dependency grammar
- Identify tasks to which syntactic structure is relevant
- Implement constituency parsing using the CYK algorithm
- Implement dependency parsing using transition-based parsing

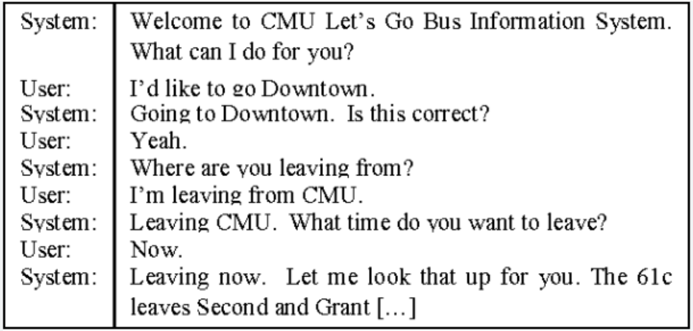


Module 8: Semantics

At the end of this module, students will be able to:

- Characterize the meanings of words in relation to one another using wordnets
- Articulate the desirable qualities of meaning representation languages
- Use syntactic parses to build a semantic role labeling system

Module 9: Discourse

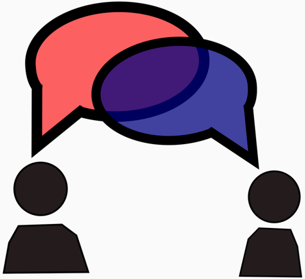


System:	Welcome to CMU Let's Go Bus Information System. What can I do for you?
User:	I'd like to go Downtown.
System:	Going to Downtown. Is this correct?
User:	Yeah.
System:	Where are you leaving from?
User:	I'm leaving from CMU.
System:	Leaving CMU. What time do you want to leave?
User:	Now.
System:	Leaving now. Let me look that up for you. The 61c leaves Second and Grant […]

Table 1. *Example of a dialog in Let's Go SDS*

http://www.cs.cmu.edu/ gparent/papers/Parent-Eskenazi-Entrainment-Interspeech2010-approved-final.pdf

Module 9: Discourse



At the end of this module, students will be able to:

- Explain how coreference resolution and entity linking work
- Articulate how discourse cohesion can be defined and quantified
- Characterize discourse coherence
- Describe the qualities that differentiate dialogue from other kinds of discourse (from the perspective of NLP)

Module 10: Speech

You are driving through Downtown Pittsburgh. You feel a sudden need to listen to listen to “Levitating” by Dua Lipa so you say “Hey Siri, play ‘Levitating’ on Apple Music.” You are using AUTOMATIC SPEECH RECOGNITION or ASR.

Siri answers back, “Playing ‘Levitating’ by Dua Lipa featuring Da Baby on Apple Music.” Siri is using SPEECH SYNTHESIS or TEXT-TO-SPEECH (TTS).

Module 10: Speech

At the end of this module, students will be able to:

- Identify ways in which processing speech is different from processing written language
- Describe appropriate approaches to speech recognition and speech synthesis

Resources

Lectures

Lectures will cover the fundamentals of the course content. To fully benefit from them, you will need to do the readings and the homework assignments (including the optional ones).

- **Slides** will be provided in advance of each lecture for note-taking purposes
- In-person students are expected to **attend each lecture**
- Remote students will be provided with a **Zoom link** for lectures; this will not be available to students in Pittsburgh except in cases of illness, etc.
- Recordings of the lectures will be made available within 24 hours of the lecture

Textbook (Free)

- For years, the standard NLP textbook has been Jurafsky and Martin's "Speech and Language Processing," Second Edition (SLP2)
- It is very outdated, but Jurafsky and Martin have been working on a revision for several years
- SLP3 is almost done, and the draft is posted on Jurafsky's website at `https://web.stanford.edu/~jurafsky/slp3/`
- SLP3 will be our sole textbook and all readings will be drawn from it
- Some content relevant to assignments and exams will be found in the textbook but not the lectures

THE TEXTBOOK IS FREE!

Office Hours

- Each TA will hold two office hours (one hour each) each week
- The instructors will schedule one office hour per week
- Hours are to be announced this week
- You are strongly encouraged to attend office hours frequently

Infrastructure

- Canvas: `https://canvas.cmu.edu`
  - Submit project milestones
  - Download slides and access videos
  - Check your grade
- Gradescope: `https://www.gradescope.com`
  - Submit homework assignments
  - Complete exams
- Piazza: `https://piazza.com`
  - Receive course announcements (very important)
  - Ask questions and read answers

Assessments

Overview of Assessments

ASSESSMENT	POINTS	PERCENTAGE
Homework	18	15.00%
Midterm Exam	20	16.67%
Final Exam	30	25.00%
Project	52	43.33%
	120	**100.00%**

Project

The largest component of the course will be a semester-long group project: question generation and question answering

- Groups of four, self-selected
- Open-ended (teams may chose from a wide variety of techniques)
- Preliminary evaluation by teaching staff
- Competitive evaluation by students
- Progress report (video)
- Final Report (technical paper)

Project Components

COMPONENT	POINTS	PERCENTAGE	DUE DATE
Literature Search	2	3.85%	Jan 26
Proposal	4	7.69%	Feb 07
Progress Report	7	13.46%	Feb 21
QA Prototype	8	15.38%	Mar 21
QG Prototype	8	15.38%	Mar 30
Final System	11	21.15%	Apr 18
Final Report	12	23.08%	Apr 28
Project Total	**52**	**100.00%**	

Homework Assignments ($\times$ 9)

- There will be nine short homework assignments
- Most will be coding assignments (autograded)
- They are designed to be completed in approximately four hours

Exams

There will be two exams, consisting of long, multipart problems (typically 3–5 questions).

- **Midterm Exam:** Exam covering Module 1–5 (Mar 2). **3 questions**.
- **Final Exam:** Comprehensive exam, focusing on Modules 6–10 (scheduled by registrar). **6 questions**.

The questions will usually center around an imagined scenario in which someone needs to solve an NLP problem. Some of the questions will be focused, involving specific calculations and execution of algorithms. Others will be more conceptual.

We will not be able to accommodate students who are not physically present on their campus during the final exam. Please do not schedule travel until the final exam schedule is released.

Time Management

Weekly Task Breakdown

You are expected to spend 12 hours per week on this course. The breakdown of a typical week might be as follows:

TASK	HOURS/WEEK
Study readings and notes	2
Attend lectures	3
Complete HW assignment	3
Work on project	4
Total	**12**

We encourage you to schedule time now. If you have difficulty completing these tasks in the allotted times, please let us know.

Policies

Late Work

Life is full of unexpected events. **If you encounter such an event, and are having trouble completing work on time, contact us!** We will try to make a plan with you that allows you to finish the course work in a reasonable time frame. This may include extended deadlines.

We cannot grant you an extension retroactively except in exceptional circumstances (an adverse event occurred very close to the deadline). However, we are happy to extend deadlines for **deaths and funerals (whether yours or those of family members), illnesses, mental health crises or episodes, weddings, important religious and national holidays, job interviews, AND SO FORTH**.

Disability

Many people have disabilities, including members of our own families. **We see disabilities as deficits not in disabled people but in the institutions and societies that are structured such that disabled people are disadvantaged**. We wish to do our part to overcome this disparate treatment. If you have a disability (visible or invisible), please let us know as soon as possible (you don't need to tell us the nature of the disability) and work with Disability Services to develop a set of accommodations which we can then approve. These may include extra time on exams, a quiet place in which to take an exam, alt text on all images, documents that work for people with differences in vision, sign language interpretation, captioning, etc.

Diversity, Equity, and Inclusion

Throughout human history, some people have been denied the rights and opportunities available to others on the basis of their race, gender, economic class, caste, ancestry, language community, age, religion, beliefs, political affiliation, and abilities (visible and invisible). A single course cannot undo the injustices of history, but we—as a teaching staff—are committed to fighting inequity and promoting inclusion. We encourage you to join us. If you feel that you, or those around you, have been treated unfairly based upon their identity (or perceived identity) by us, by other members of the teaching staff, or by other students in the course, we ask that you share your experience with the Center for Student Diversity and Inclusions (csdi@andrew.cmu.edu, (412) 268-2150) or by using the anonymous Report-It reporting platform (reportit.net, username: `tartans`, password: `plaid`).

Questions?