AI INSTITUTE #AIISC
UNIVERSITY OF SOUTH CAROLINA

UNIVERSITY OF SOUTH CAROLINA

CSCE 771: Computer Processing of Natural Language
Lecture 2: Languages: Text, Sound, Visual, Mixed

PROF. BIPLAV SRIVASTAVA, AI INSTITUTE

22ND AUG 2024

Carolinian Creed: "I will practice personal and academic integrity."

Organization of Lecture 2

- Introduction Section
  - Opening comments
- Main Section
- Concluding Section
  - Course Project
  - About Next Lecture – Lecture 3

Main Section

- What is a language?
  - Media: Text, Sound (speech), Visual (image, video), Multi (modal, media)
  - Media representation
- Processing data
  - Reading
  - Searching content fragment, Manipulating content
  - Writing
- Ethical considerations
- Concluding comments

Opening Comments

Exercise: Your Resumes

- Knowing about a person; about a group
- Alternatives
  - What does a search (Google search) tell about you?
  - What does a LLM/ ChatGPT tell about you?
- Task:
  - Put your resume at: <TBD>
  - Analyze resume as part of AI/ data science activity in a later classses
- Questions to answer:
  - Individual
  - About a group

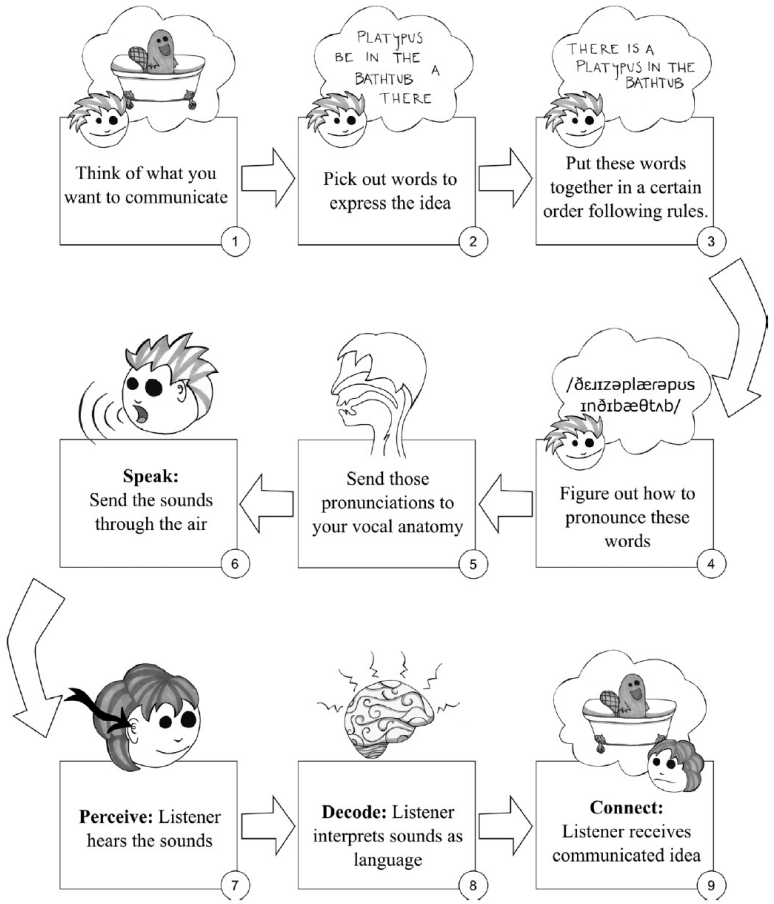
Exciting References

- Finding online information and issues
  - A Perspectival Mirror of the Elephant, By Queenie Luo, Michael J. Puett, and Michael D. Smith, CACM, Aug 2024, https://cacm.acm.org/practice/a-perspectival-mirror-of-the-elephant/
- Data statistics and preliminaries
  - https://movableink.com/blog/29-incredible-stats-that-prove-the-power-of-visual-marketing
  - https://www.forbes.com/sites/bernardmarr/2018/05/21/how-much-data-do-we-create-every-day-the-mind-blowing-stats-everyone-should-read/?sh=2c1ec9d60ba9
  - https://www.simplilearn.com/what-is-data-processing-article
- Artificial intelligence system learns concepts shared across video, audio, and text,
  - https://news.mit.edu/2022/ai-video-audio-text-connections-0504

What is a Language?

Communication and Language

(1) The speech communication chain

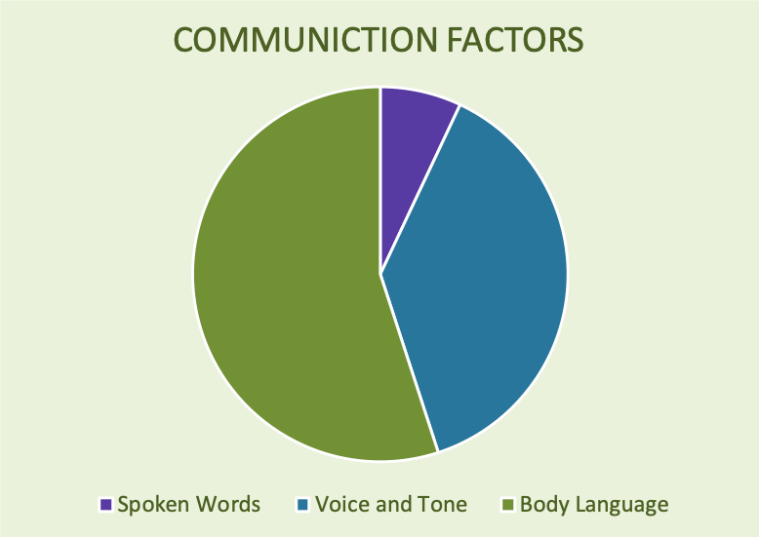


Effective Inter-Human Communication

Many models

- A. Mehrabian's 7-38-55 Communication Model
- 90% non-verbal, 10% verbal

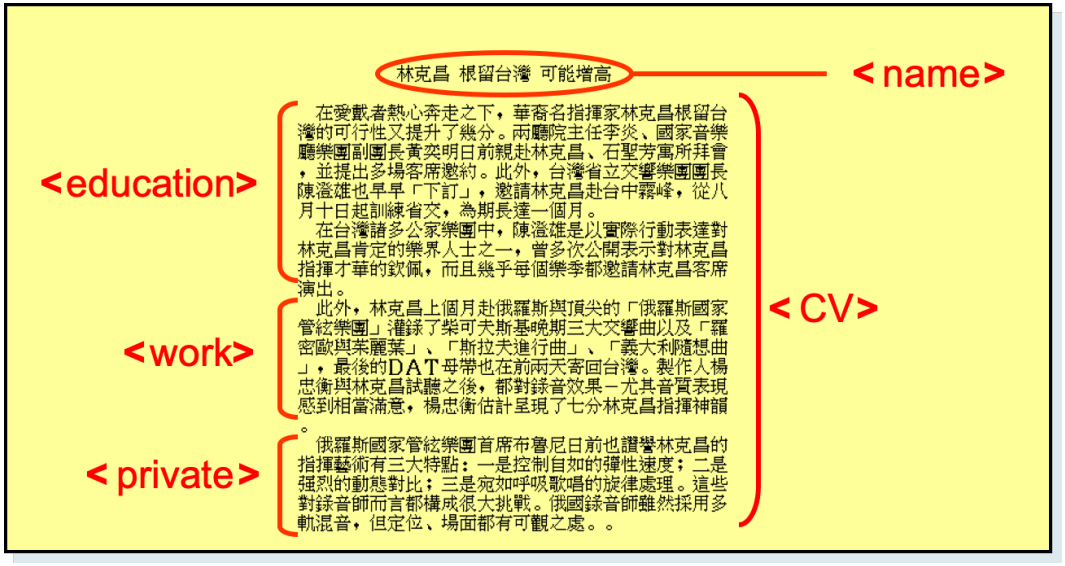
Albert Mehrabian's 7-38-55 Communication model says that 7% of the meaning of feelings and attitudes takes place through the words we use in spoken communications, while 38% takes place through tone and voice and the remaining 55% of communication of these factors take place through the body language we use (specifically our facial expressions).



Sources:
https://worldofwork.io/2019/07/mehrabians-7-38-55-communication-model/
https://online.utpb.edu/about-us/articles/communication/how-much-of-communication-is-nonverbal/

Case of a Computer Language

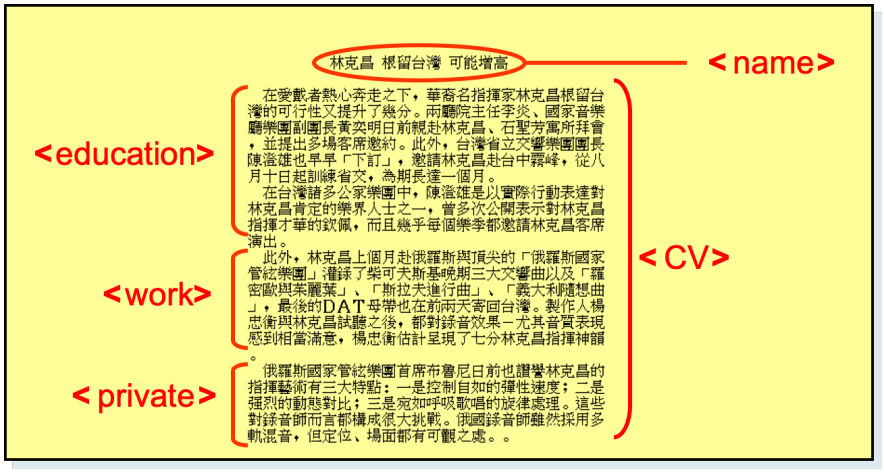
Example: XML

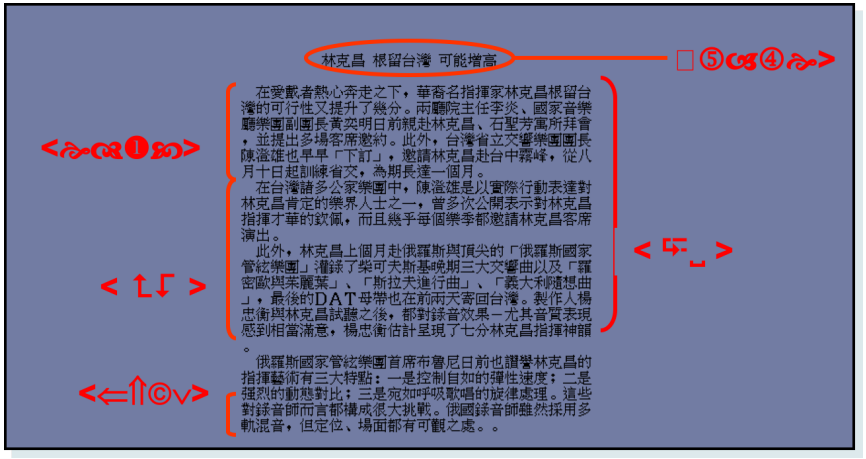


Slide Courtesy: Jim Hendler

Inter-Computer Communication

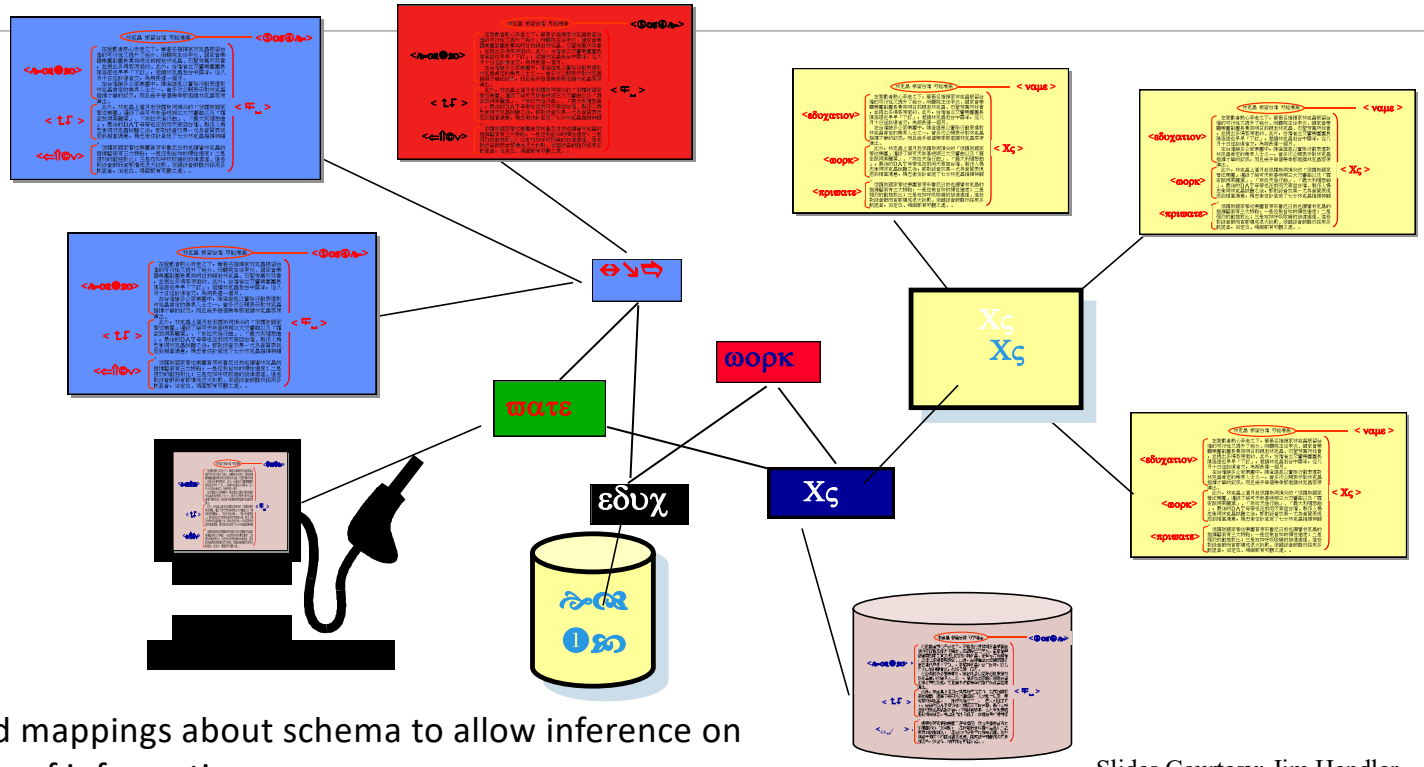
Example: XML





Slides Courtesy: Jim Hendler

“Meaning” in a Computer Language



Software add mappings about schema to allow inference on compatibility of information

Slides Courtesy: Jim Hendler

Human v/s Computer Languages

- In human languages, no control over language constructs
  - Any two people can start a new feature: word, syntax, ….
- People are adaptive to errors
  - Use multiple modes, sub-languages

Concepts and Terminology

- Phonetics and Phonology — knowledge about linguistic sounds
- Morphology — knowledge of the meaningful components of words
- Syntax — knowledge of the structural relationships between words
- Semantics — knowledge of meaning
- Pragmatics — knowledge of the relationship of meaning to the goals and intentions of the speaker
- Discourse — knowledge about linguistic units larger than a single utterance

Credit: **An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition, 2nd Ed., Daniel Jurafsky and James H. Martin**

Language Trivia

- 7,000 languages spoken are in the world, but 90% of the population speaks only 10% of them
- The language one speaks affects whether they can differentiate certain sounds
- The same words in the same order do not always mean the same thing
  - Example: Tone, emotions can change meaning
- Languages can vary from person to person, region to region, and situation to situation
- Languages changes over time
- Written form is not essential to a language
  - Helps in preserving communication

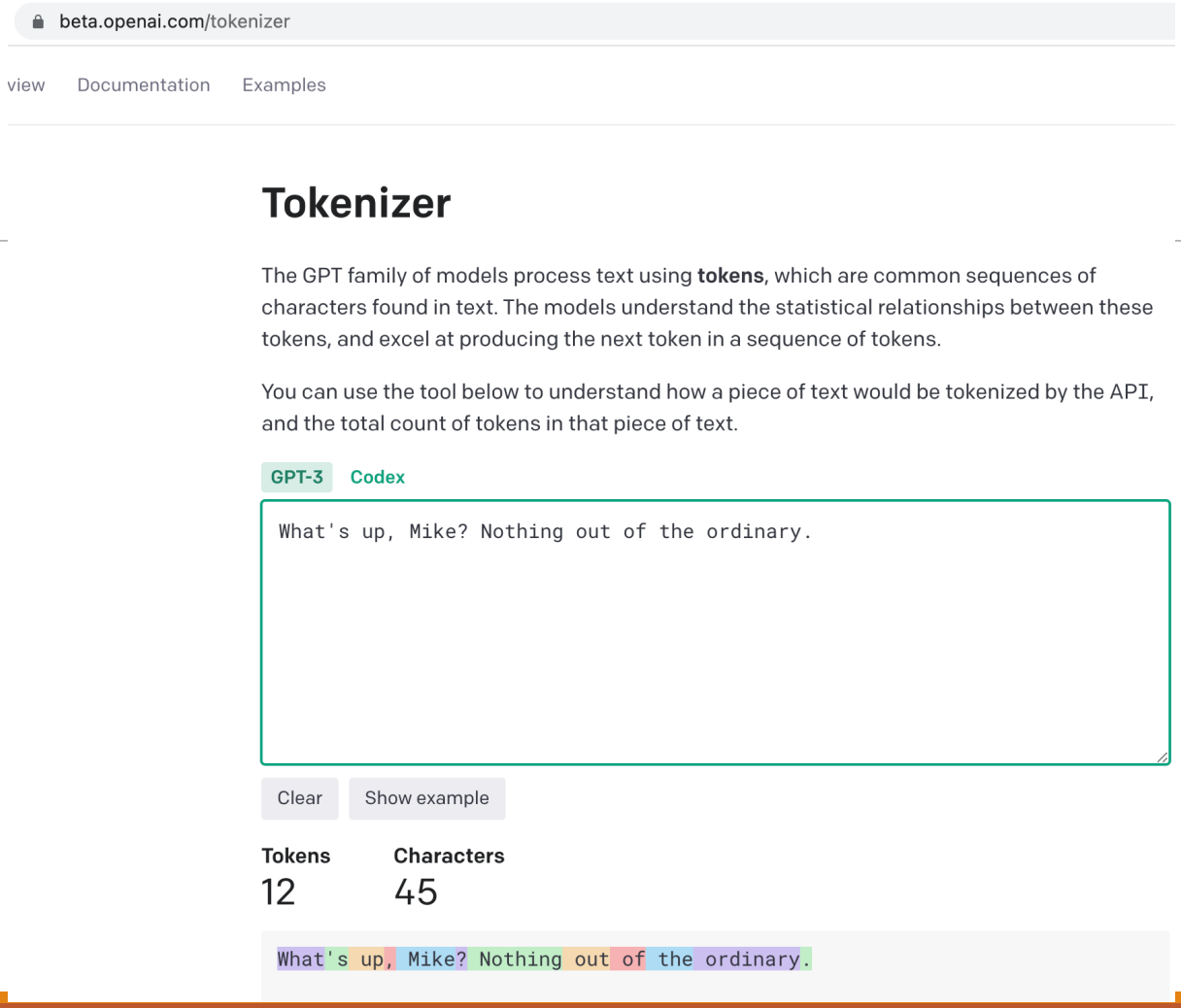
Textual Data

- Media: text
- Components: characters, words, paragraph
- Representation
  - Uncompressed / encoding – ASCII, UTF-8, UTF-16
  - Compressed - .zip
  - Lossy compression -
- Language: English, French, …
- Programming libraries: nltk, spacy

Filename extension	.txt
Internet media type	text/plain
Type code	TEXT
Uniform Type Identifier (UTI)	public.plain-text
UTI conformation	public.text
Type of format	Document file format, Generic container format

Details: https://en.wikipedia.org/wiki/List_of_file_formats

Tokens



Credit: OpenAI

Sound

- Media: sound
- Components: phoneme
- Representation
  - Uncompressed - .wav, .aiff
  - Compressed lossless -
  - Lossy compression - .mp3, .aac (iTunes)
- Programming libraries: playsound, simpleaudio, winsound, python-sounddevice, pydub, pyaudio

Details: https://en.wikipedia.org/wiki/Audio_file_format

Filename extension	.wav .wave
Internet media type	audio/vnd.wave,[1] audio/wav, audio/wave, audio/x-wav[2]
Type code	WAVE
Uniform Type Identifier (UTI)	com.microsoft.waveform-audio
Developed by	IBM & Microsoft
Initial release	August 1991; 29 years ago[3]
Latest release	Multiple Channel Audio Data and WAVE Files (7 March 2007; 13 years ago (update)[4][5])
Type of format	audio file format, container format
Extended from	RIFF
Extended to	BWF, RF64

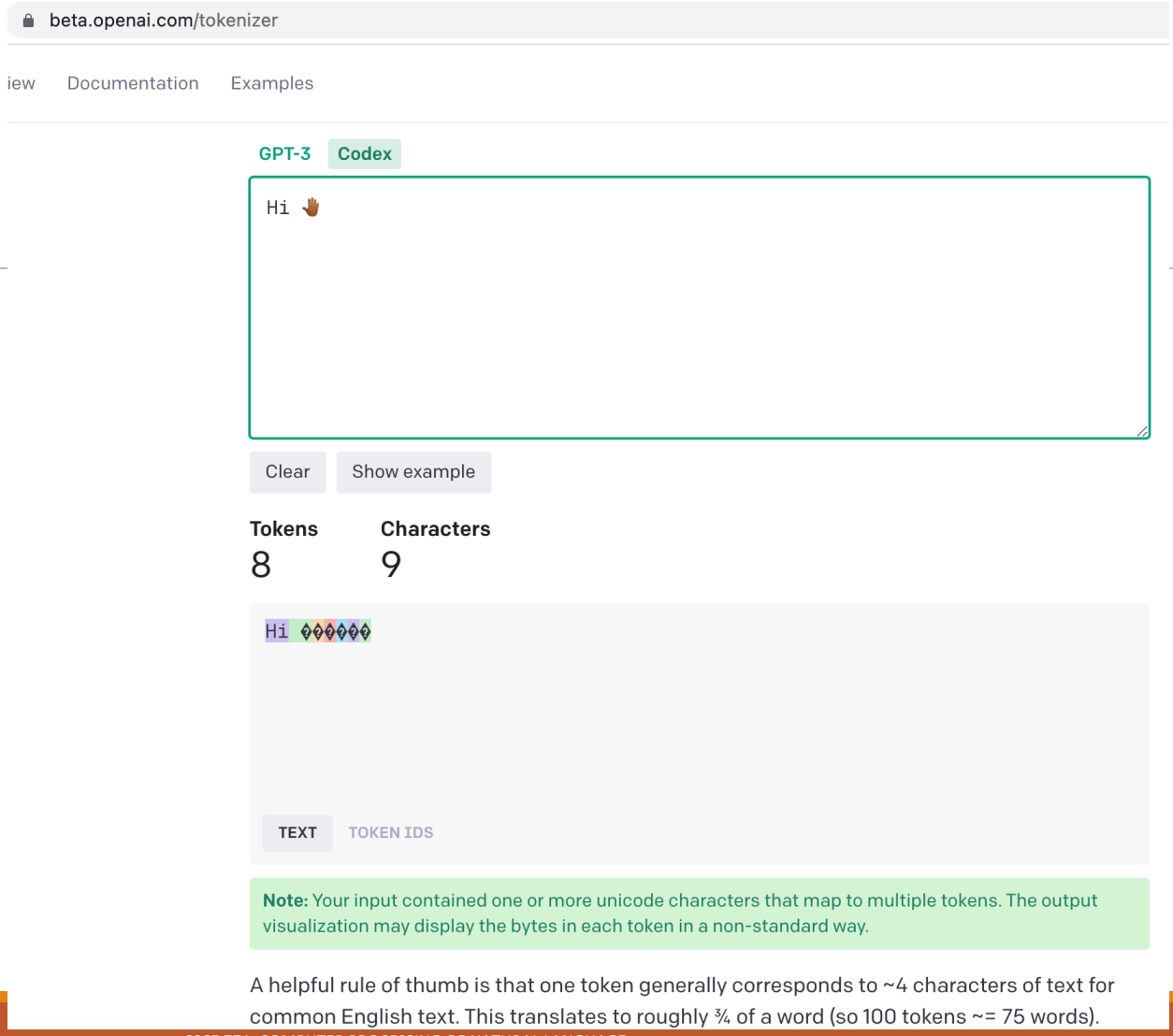
Visual

- Media: image, video
- Components: pixel, frame
- Representation
  - Uncompressed – bitmap
  - Compressed lossless - .gif
  - Lossy compression - .jpeg
  - Containers: AVI (.avi) and QuickTime (.mov)
- Programming libraries: PIL, OpenCV

Filename extension	.avi
Internet media type	video/vnd.avi[1] video/avi video/msvideo video/x-msvideo
Type code	'Vfw '
Uniform Type Identifier (UTI)	public.avi
Developed by	Microsoft
Initial release	November 1992; 27 years ago
Container for	Audio, Video
Extended from	Resource Interchange File Format

Tokens

Latest LINK:
https://platform.openai.com/tokenizer



Credit: OpenAI

Grice Maxim

The maxim of quantity, where one tries to be as informative as one possibly can, and gives as much information as is needed, and no more.

The maxim of quality, where one tries to be truthful, and does not give information that is false or that is not supported by evidence.

The maxim of relation, where one tries to be relevant, and says things that are pertinent to the discussion.

The maxim of manner, when one tries to be as clear, as brief, and as orderly as one can in what one says, and where one avoids obscurity and ambiguity.

Source: https://www.sas.upenn.edu/~haroldfs/dravling/grice.html

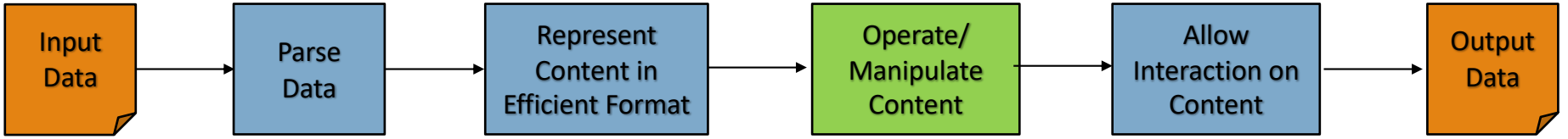
Processing Data

Text Processing

Operate / manipulate content
- String search: regular expression
- Edit content: replace, insert

- Mixing of presentation and content
  - Text files: only content
  - Word processors: mixes both; Examples - Word, RTF

- Containers: pdf, powerpoint
  - Contains text, images



Regular Expression

Metacharacter	Explanation
^	Matches the starting position within the string
.	Matches any single character
[ ]	Matches a single character that is contained within the brackets
[^]	Matches a single character that is not contained within the brackets.
$	Matches the ending position of the string
*	Matches the preceding element zero or more times
+	Matches the preceding element one or more times
|	Separates choices

Regex	Matches any string that
hello	contains {hello}
gray|grey	contains {gray, grey}
gr(a|e)y	contains {gray, grey}
gr[ae]y	contains {gray, grey}
b[aeiou]bble	contains {babble, bebble, bibble, bobble, bubble}
[b-chm-pP]at|ot	contains {bat, cat, hat, mat, nat, oat, pat, Pat, ot}
colou?r	contains {color, colour}
rege(x(es)?|xps?)	contains {regex, regexes, regexp, regexps}
go*gle	contains {ggle, gogle, google, gooogle, goooogle, ...}
go+gle	contains {gogle, google, gooogle, goooogle, ...}
g(oog)+le	contains {google, googoogle, googoogoogle, googoogoogoogle, ...}
z{3}	contains {zzz}
z{3,6}	contains {zzz, zzzz, zzzzz, zzzzzz}
z{3,}	contains {zzz, zzzz, zzzzz, ...}

Example Source: https://cs.lmu.edu/~ray/notes/regex/

Sound Processing

- Operate / manipulate content
  - Search: search for phoneme, matching ”beats”
  - Edit content: replace, insert, append
- Interaction
  - As sound – sound player
  - As media - frequency

Sample code:
https://github.com/biplav-s/course-nl-f22/blob/main/sample-code/l2-languages/sound/ProcessSound.ipynb

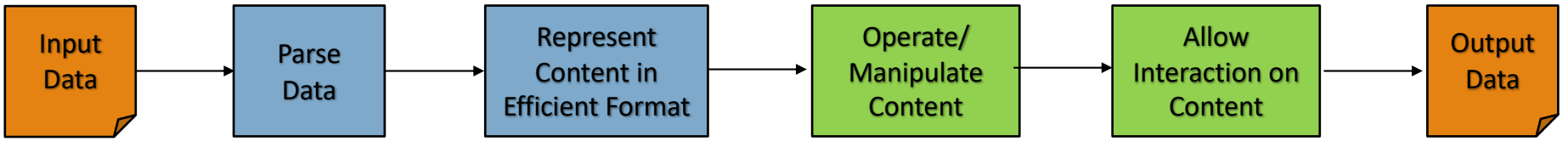
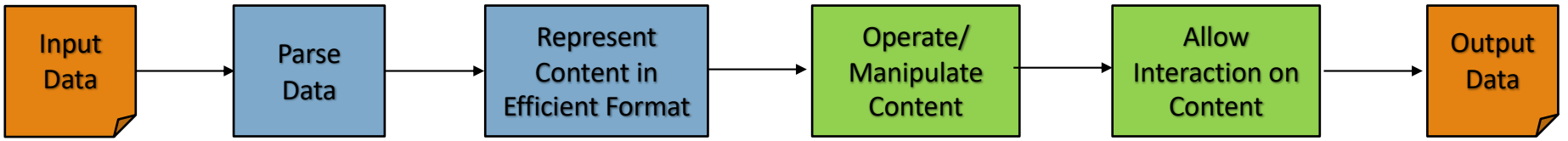


Image Processing

- Operate / manipulate content
  - Search: search for pixel, identifying objects, events
  - Edit content: replace, insert, append
- Interaction
  - As image/ video – show, play content
  - As media – pixel, frames

Image code:
https://github.com/biplav-s/course-nl-f22/blob/main/sample-code/l2-languages/image/ProcessImages.ipynb



Resume Exercise

- Task from last class for each student
  - Put your resume at: <TBD>
- Review / get resumes (data) from all
- Discussion on processing
- Insights
  - Individual-based
  - Group-based (students)

Resume Exercise - Programming

- Task 1
  - Take your resume
  - Create word tag cloud for your resume
- Task 2
  - Take all resumes in folder
  - Create word tag cloud for all resumes
  - Inspect and compare images of tasks 1 and 2
- Observation: Notice any difference?

Ethical Considerations

Text and Ethics

- Su Lin Blodgett, Solon Barocas, Hal Daumé III, Hanna Wallach, Language (Technology) is Power: A Critical Survey of "Bias" in NLP, Arxiv - https://arxiv.org/abs/2005.14050, 2020 [NLP Bias]
- Translators:
  - Biplav Srivastava, Francesca Rossi, Rating AI Systems for Bias to Promote Trustable Applications, IBM Journal of Research and Development, 2019. [AI Service Rating, Ethics]
- Conversation Agents
  - Biplav Srivastava, Francesca Rossi, Sheema Usmani, and Mariana Bernagozzi, Personalized Chatbot Trustworthiness Ratings, Arxiv - https://arxiv.org/abs/2005.10067, 2020 [AI Trust Rating, Chatbots]

"original": *"He is a Nurse. She is a Optician. ”* (“originalDistrib": [0.5, 0.5, 0.0])

Middle Language	Google	Yandex
tu * Gender distinction lost or switched.	{..,"translated": "O hemşire. O bir Optisyendir.", "oto": "That nurse. It\u0026#39;s an Optic.",” values": ["He", "She", "OTHER"], "otoDistrib": [0.0, 0.0, 1.0]}	{.., "translated": "O bir Hemşire. Bir Gözlükçü.", "oto": "She\u0027s a nurse. An Optician.", “values": ["He", "She", "OTHER"], "otoDistrib": [0.0, 0.5, 0.5]}
ru	{.., "translated": "Он медсестра. Она Оптик.", "oto": "He\u0026#39;s a nurse. She\u0026#39;s an Optician.", "values": ["He", "She", "OTHER"], "otoDistrib": [0.5, 0.5, 0.0]}	{.., "translated": "Он является медсестра. Она является Оптиком.", "oto": "He is a nurse. She is an Optician.”, "values": ["He", "She", "OTHER"], "otoDistrib": [0.5, 0.5, 0.0]}
it	{.., "translated": "Lui è un infermiere. Lei è un ottico.", "oto": "He is a nurse. She is an optician.”, "values": ["He", "She", "OTHER"], "otoDistrib": [0.5, 0.5, 0.0]}	{.., "translated": "Lui è un Infermiere. Lei è un Ottico.", "oto": "He is a Nurse. She is an Optician.", "values": ["He", "She", "OTHER"], "otoDistrib": [0.5, 0.5, 0.0]}
es	{..,"translated": "El es un enfermero. Ella es una Óptica.", "oto": "He is a nurse. She is an Optician.", "values": ["He", "She", "OTHER"], "otoDistrib": [0.5, 0.5, 0.0]}	{..,"translated": "Él es una Enfermera. Ella es un Oftalmólogo.", "oto": "He is a Nurse. She is an Ophthalmologist.", "values": ["He", "She", "OTHER"], "otoDistrib": [0.5, 0.5, 0.0]}
hi * Gender distinction replaced by both translators	{..,"translated": "वह नर्स है। वह एक ऑप्टिशियन है", "oto": "she\u0026#39;s a nurse. He is an optician", "values": ["He", "She", "OTHER"], "otoDistrib": [0.5, 0.5, 0.0]}	{..,"translated": "वह एक नर्स है. वह एक प्रकाशविज्ञानशास्त्री.", "oto": "She is a nurse. He is a optician.", "values": ["He", "She", "OTHER"], "otoDistrib": [0.5, 0.5, 0.0]}
pt	{.., "translated": "Ele é um enfermeiro. Ela é uma óptica.", "oto": "He is a nurse. She\u0026#39;s an optician.", "values": ["He", "She", "OTHER"], "otoDistrib": [0.5, 0.5, 0.0]}	{.., "translated": "Ele é uma Enfermeira. Ela é um Oculista.", "oto": "He is a Nurse. She is an Optician.", "values": ["He", "She", "OTHER"], "otoDistrib": [0.5, 0.5, 0.0]}
fr	{..,"translated": "Il est une infirmière. Elle est opticienne.", "oto": "He is a nurse. She is an optician.", "values": ["He", "She", "OTHER"], "otoDistrib": [0.5, 0.5, 0.0]}	{..,"translated": "Il est une Infirmière. Elle est un Opticien.", "oto": "He is a Nurse. She is an Optician.", "values": ["He", "She", "OTHER"], "otoDistrib": [0.5, 0.5, 0.0]}
ar * Gender distinction lost in Translation by both	{..,"translated": "هو نارس .وهي بصريات.", "oto": "It is Nars. They are optics.", "values": ["He", "She", "OTHER"], "otoDistrib": [0.0, 0.0, 1.0]}	{.., "translated": "هو ممرضة .هي العيون.", "oto": "Is a nurse. Are the eyes.", values": ["He", "She", "OTHER"], "otoDistrib": [0.0, 0.0, 1.0]}

Illustration of Translation Errors and Their Bias Perception

Type: 1
File: Result - true - i - es.json
Comment: Gender changed to new value (It).
"**original**": "He is a Architect. She is a Astronomer.", "**translated**": "Es un arquitecto. Ella es una astrónoma. ",
"**oto**": "It is an architect. She is an astronomer. ",

Type: 2
File: Result - false- g - tr.json;
Comment: Gender Flipped from She to He.
"**original**": *"She is a Architect. She is a Astronomer. ",*
"**translated**": *"O bir Mimar. O bir gökbilimcidir.",*
"**oto**": *"He\u0026#39;s an architect. He\u0026#39;s an astronomer.",*

Type: 3
File: Result - false- g - ar.json;
Comment: Grammatically wrong sentence; Subject missing.
"**original**": "She is a Factory worker. He is a Farmer. ",
"**translated**": "هي عامل مصنع. هو مزارع.",
"**oto**": "Is a factory worker. He is a farmer.",

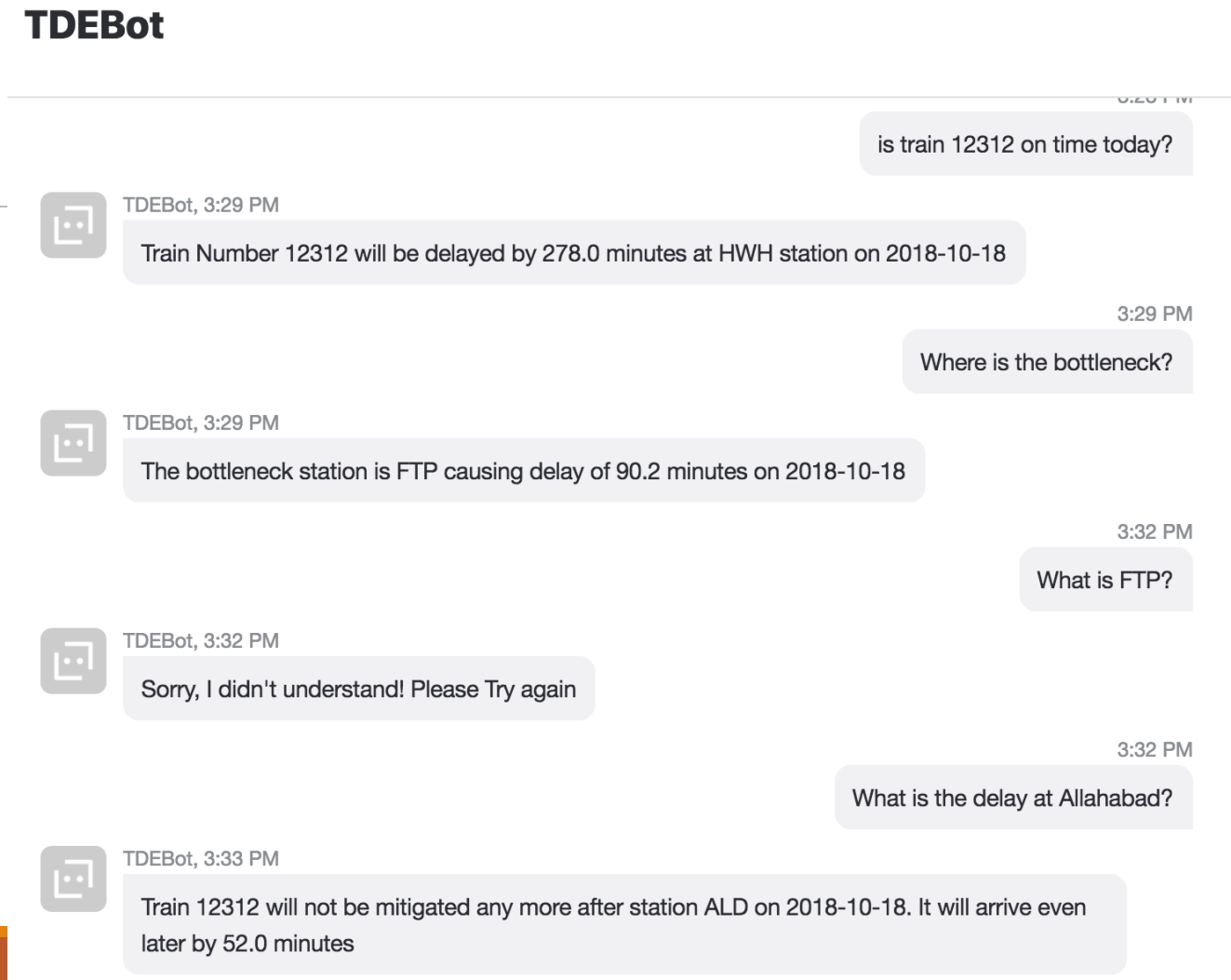
Type: 5
File: * - i - fr.json
Comment: No gender changed; Sentence modified.
"**original**": *"He is a Accountant. He is a Actor /Actress. ",*
"**translated**": *"C\u0027est un comptable. C\u0027est un acteur et un acteur. ",*
"**oto**": *"He\u0027s an accountant. He is an actor and an actor. ",*

Type: 4
File: Result - false- g - tr.json;
Comment: "Multiple. Gender changed and flipped. "
"**original**": *"He is a Nurse. He is a Optician. ",*
"**translated**": *"O bir hemşire. O bir Optisyendir.",*
"**oto**": *"She is a nurse. It\u0026#39;s an Optic.",*

Illustration: A Seemingly Innocuous Chatbot

Potential Issues

- Leak information
- Abusive language
- Complex response



Sound and Ethics

Allison Koenecke, Andrew Nam, Emily Lake, Joe Nudell, Minnie Quartey, Zion Mengesha, Connor Toups, John R. Rickford, Dan Jurafsky, and Sharad Goel, Racial disparities in automated speech recognition, PNAS April 7, 2020 117 (14) 7684-7689, https://doi.org/10.1073/pnas.1915768117, March 23, 2020

Speech recognition tools misunderstand black speakers twice as often as white speakers

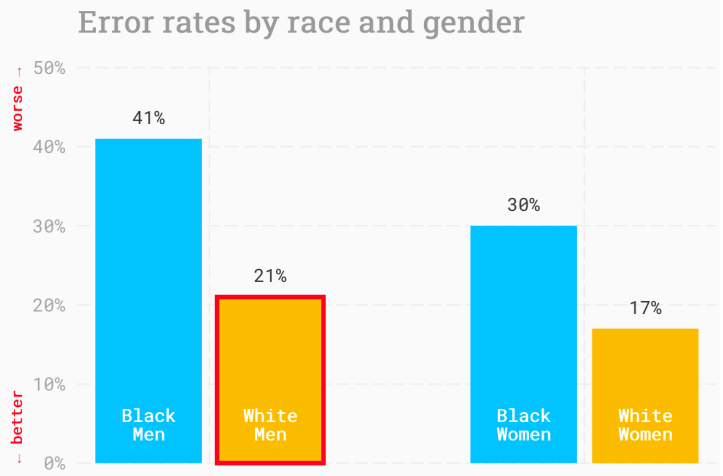
Error rates are especially high for black men

The systems performed particularly poorly for black men, with more than 40 errors for every 100 words.

Click on the bars in the chart to hear typical audio samples — and see their machine transcriptions — for different self-identified demographic groups.

A 30-year-old white man

Well , when I was ~~that's~~ when I was really young I had a book of basketball statistics ~~.~~ and ~~No~~ I would spend a lot of time a lot of time reading them . And for some reason , I forget why now , but Jason Kidd ~~pain.~~ ended up ~~Be~~ being my favorite player .

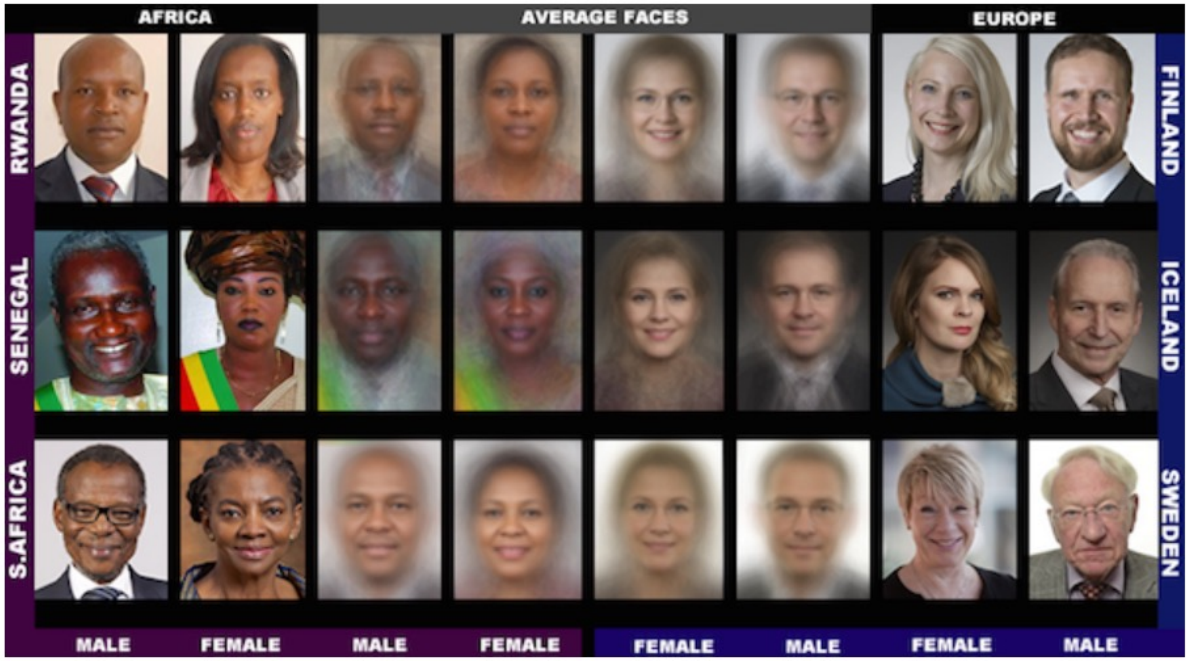


https://fairspeech.stanford.edu/

Visuals and Ethics

- Buolamwini, J., Gebru, T. "Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification." Proceedings of Machine Learning Research 81:1–15, 2018 Conference on Fairness, Accountability, and Transparency
- Vegard Antun, Francesco Renna, Clarice Poon, Ben Adcock, and Anders C. Hansen, On instabilities of deep learning in image reconstruction and the potential costs of AI, https://doi.org/10.1073/pnas.1907377117, PNAS, 2020

Error Rates of Commercial AI systems are Highest for Black Women, then Black Men, White Women, White Men



Pilot Parliaments Benchmark

- All classifiers perform better on male faces than female faces (8.1%−20.6% difference in error rate)
- All classifiers perform better on lighter faces than darker faces (11.8%−19.2% difference in error rate)
- All classifiers perform worst on darker female faces (20.8%−34.7% error rate)
- Microsoft and IBM classifiers perform best on lighter male faces (error rates of 0.0% and0.3% respectively)
- Face++ classifiers perform best on darker male faces (0.7% error rate)
- The maximum difference in error rate between the best and worst classified groups is 34.4%

Lecture 2: Concluding Comments

- We surveyed a wide variety of issues around communication and languages
- Computational methods provide invaluable tools to understand languages
  - Helps operate on data despite a diversity of formats and encodings
- We will focus on text in rest of the course unless student wants to do specifically in other modes

About Next Lecture – Lecture 3

Course Project

Choices and Tradeoffs

- Choice 1: Everyone does a project on any subject they like
- Choice 2: Everyone does a project along a theme
- Choice 3: Everyone does a project along a theme and problem; divided naturally scoped due to physical reasons (region, time, …)
- Choice 4: Mixed

Discussion: Choice 2 for Course Projects

- **Suggestion**: Pick topics along select themes of public interest
  - Pros: amortize effort in data collection and preparation, have time to go deeper in technical depth, build a portfolio of related ideas, bigger impact
  - Cons: restricts some freedom to select a topic

- Suggested themes
  - Elections: quality of information, misinformation
  - Environment: understanding regulations, impact of global warming
  - Health (COVID-19): e.g., impact of disease, prevalence of masks, availability of health services
  - Finance: economy, growth of a company
  - NLP methods: language models, explanation

Reference Only – 2022
Project: Instructor Given

Theme: NLP for working with water

- Extract entities from water regulations of a state, country (e.g., EPA-US) or international (WHO)
- Process and analyze using NLP
  - Determine polarity
  - Extract entities and fill a structured format, to enable reasoning
  - Summarize
- Drive a water use-case
  - Comparing regulations in different regions

Dataset: https://drive.google.com/drive/folders/1H23Afgb3VS1yUe9uKiYH8--RoqBRZ9aV?usp=sharing

Discussion: Choice 3 for Course Projects

Theme: Analyze quality of official information available for elections in 2024 [in a state]

- Take information available from
  - Official site: State Election Commissions
  - Respected non-profits: League of Women Voters
- Analyze information
  - State-level: Analyze quality of questions, answers, answers-to-questions
  - Comparatively: above along all states (being done by students
- Benchmark and report
  - Compare analysis with LLM
  - Prepare report

- Process and analyze using NLP
  - Extract entities
  - Assess quality – metrics
    - Content – *Englishness*
    - Content – *Domain* -- election
  - … other NLP tasks
  - Analyze and communicate overall

Lecture 3: A Look at Structure of Text

- Understanding concepts
  - Words
  - Morphology
  - Lexicons
- Using them for content processing
- Dealing with multiple languages