1.1 Outline

- kurzfristige Umstellung auf Zoom
- Diskussion Survey Seminarerwartungen
  - 3/4 mit R-Erfahrung, wenige mit Python und Kommandozeile
  - SoCom Leute mit Inhaltsanalyse, keine Daten in Aussicht
  - Einführungskurs, aber komplementäres Wissen und Pointers für Fortgeschrittene
  - Skills für Seminar/BA-Arbeiten
  - allgemeine Programmier- und Computerkenntnisse
- Fragen zu Inhalt/Website?
- Diskussion letzte Sitzung + Paper
- Hauptteil: Bedeutung/Grundlage von Textanalyse
  - Auf welcher methodischen Grundlage steht das Feld?
  - Qualitativer Anteil gegenüber letzten Sitzung herausheben
- zweite Lektion: erste Textanalyse
  - einfach, aber mächtig

1.2 Recap last Lecture

- Technologie verändert Welt. Seit immer, erneut grosser Schritt nach Industrialisierung.
- Computer als Werkzeug/interaktiver Partner
  1. nehmen an sozialen Prozessen teil -> wie verändert sich das Soziale/Ökonomische (Wissenschaft, Arbeit, Jobmarkt, Benachteiligung)?
  2. für CSS: Daten wichtiger als ML
- Daten sind da -> erst Programmieren ermöglicht Auswertung

1.3 Reading

- hochaktuell: Einsichten in Pandemiegeschehen durch Netzwerkanalyse
- methodical focus because of Nature paper
- more than self-reported data (survey)
- tlw. schwieriger Zugang
  - organisationsintern Daten und Datenschutz
  - historische und textuelle Daten einfacher zugänglich

1.4 Semiotic Triangle

- Was ist Sprache?
  - Keine Philosophie-Vorlesung
  - technisch auch von Bedeutung
- Versuch der Einheit: Ding, Konzept und Wort
- keine direkte Beziehung zwischen Symbol & Gegenstand

  - keine Eineindeutigkeit wie in Datenbanken –> schwierig für Computer
    * identische Personen- und Ortsnamen
    * umfasst Früchte auch Hülsenfrüchte?
  - jede Ecke kann wechseln
    * Gleiches heisst anders, anderes heisst gleich
- zweiteilige These umstritten (Sapir-Whorf-Hypothese)
  - Sprache formt das Denken
  - unabhängig der Determination: überragende Bedeutung für das Soziale
  - Inuit-Anekdote zu Schnee bedingt durch Grammatik
- Sprache ist das Soziale schlechthin
  - Vermittlungsmedium
  - weitere Formen: Zeichnen, Mathematik, Fotos
- wenn nicht kommuniziert, dann gesellschaftlich ohne Bedeutung (aber nicht unbedingt unvorstellbar)
  - Wörter sind Unterscheidungen
  - Link zu Luhmann
- Aktuelles Beispiel Ukraine-Krieg
  - Konflikt vs Krieg (Gewalt) vs Invasion (asymmetrisch), militärische Operation
  - Definitionskampf ist gut erkennbar von Russland, aber auch allen anderen

2.4 Working with Texts
2.5 A micro and macro perspective I

- Nun klar, wieso Textanalyse wichtig, aber welche Herangehensweise?
- Traditionell
  - Inhaltsanalyse, close reading
  - Einzeldokumente
  - lange Zeit alternativlos
- computergestützte Textanalyse
  - NLP, distant reading
  - Textsammlungen
- Rauszoomen bringt mehr/neues Verständnis, nicht nur Reinzoomen
- Methodik ändert evtl. Fragestellung
  - NLP: nicht Individuum, sondern Diskurs/Gesellschaft/Gruppe
  - strukturelle Beschreibungen und Kultur/Stimmung

2.6 A micro and macro perspective II

- je ein Problem je Approach
  - close: nicht skalierbar
    * ist das generalisierbar?
  - distant: kontextlos, da Narrativ/Einzelheiten verloren gehen

    * verlieren wer/was/wo/wie/wann/warum
    * was bedeuten Zahlen? Verweis: BIP (informelle Wirtschaft)

2.7 From micro to macro :bar_chart:...and back again :bookmark_tabs:

- Lösung: Vogelperspektive, dann Eintauchen und zurück
- Gute Data Science besteht aus guter Kenntnis von Daten
- Grösser nicht immer besser

2.8 Two Research Paradigms

- genauere Einordnung: exaktere Epochenbestimmung
- Agnostik/Induktion ausnutzen für anderes Narrativ
  - data-driven Diskurs ordnen
- Modelvorhersagen zu Kausalitätsaussagen
  - z.B. Klimawandel Berichterstattung -> Erfolg grüne Partei?
  - Metadaten zu Kommunikationsflüsse nötig

2.9 Numbers do not talk :no_mouth:

- alter Konflikt Quali/Quanti
  - beide Lager kritisch gegenüber NLP
  - zu wenig rigoros, zu naiv mangels Kontext
- Zahlen sprechen nicht für sich selbst
- komplementär

2.10 Text as Data

- Link zu semiotischem Dreieck
- Text inhärent schwierig
  - herausfordernste Datenform
  - Front der AI
- Wörter = diskrete Symbole
  - nominales Skalenniveau
- compositional
  - grosse Mäuse, kleine Elefanten
- unstrukturiert
  - anders als Tabelle/Datenbank
  - unterschiedliche Datenformate

2.11 Unstructured Text? :thinking:
3.11 Data Formats
3.12 In-class Task: File Types

- Problem ist nicht wirklich der Text, sondern das Format

- alle möglichen Filetypen, nicht nur Text
  - zip/tar, exe, dmg/iso, jpg/png/gif
- öffnen von Editor?
- Dateiendungen aktiviert auf Computer?

3.13 File Formats

- am besten raw + open
- Papier altert langsamer als Software!
- **Pause**

4.13 Let's Dive into it! :sweat_drops:
4.14 Counting ngrams

- Google Books
  - indexiert ganze (Uni-)Bibliotheken
  - in 2009 mehr als 4% aller veröffentlichter Bücher
- See how ideas evolve/change over time
- y: relative Worthäufigkeiten
- x: Bücher indexiert nach Publikationsjahr
- publiziert in Science, kein klassiches SoWi Journal
  - disziplinare Grenzen brechen auf

4.15 In-Class Task: Investigate the Environmental Discourse

- Dauer: 20 Minuten
- issues described by whom?
- Herumgehen + selbst ausprobieren
- Wikipedia nutzen

4.16 Refine your Queries
4.17 Ngram 'pay attention'

- major shift: "call attention" -> "pay attention"
- externer Faktor (call) vs. aktives Verhalten (pay)
- pay attention as a form of currency
- Zusammenhang? Aufmerksamkeitsökonomie, Individualismus
- "if you don't want to call attention to yourself by giving an incorrect answer, then you should probably pay attention in class."

4.18 Remember :thumbsup:

- Grosse Frage ist
  - Wird das gleiche anders benannt?
  - Geht es um was anderes?

- Link zu Odgen Dreieck von nicht fixer Beziehungen

4.19 No Culturomics but Meaning-Making

- Änderung von kontextueller Verwendung oder Wortfrequenz
- Eigentum hat sich etabliert, Religion degeneriert
- Patterns EN
  - dessert=>*_ADJ
  - *=>public_ADJ
  - *=>personal_ADJ
- Pattern DE
  - Kulturen=>*_ADJ
  - Kinder=>*_ADJ
- only entire words, yet: _INF

4.20 Questions of Interpretation

- numbers don't talk
- Kommunikation
  - Weisse Schafe nicht erwähnenswert, nur schwarze
  - Nachrichtenwerte
  - Themenkonjunkturen
- einzelne Wörter bilden schlechte Evidenzbasis

4.21 A Word of Caution

- Google: ~4% of all books ever published
- Compared to the 2009 versions, the 2012 and 2019 versions:
  - more books, improved OCR, improved library and publisher metadata.
  - ngrams across page boundaries, no ngrams across sentence boundaries
  - rule-based tokenization
- wissenschaftlicher Standard
  - Ziel: nicht Unfehlbarkeit, sondern methodisch nachvollziehbar und kritisierbar
  - zitierfähig
- HathiTrust
  - curated collection
  - filter by meta data

4.22 Interacting with Data

- Lens / transformation like biology/physics
  - allerdings keine Labordaten
  - Soziales ohne ceteris paribus

- not just mapping but interacting
  - Daten erlauben neue Sicht
  - deshalb nicht CS überlassen

4.23 Prepare your System

- Nicht riskanter als anderes. Ein Backup gehört dazu, ein Datenverlust sicher nicht.
- Unklarheiten/Probleme unbedingt zurückmelden
- Wer hat Python schon installiert? Welches OS/Installer?

4.24 New room :classical_building:

5.24 Questions?

5.25 References