19.4.2024

Guidelines Assoziationen-Bereinigung

In diesem Projekt beschäftigen wir uns mit Personal Name Compounds (z.B. Willkommens-Merkel oder Tore-Klose) und den entsprechenden Namen (Angela Merkel und Miroslav Klose). Wir haben Menschen gefragt, was ihnen spontan zu den Compounds oder dem dazugehörigen Namen einfällt und sie gebeten, 3-5 Wörter oder Phrasen aufzuschreiben. Die sollten jetzt bereinigt werden, da es bei Freitexteingaben immer zu Unsauberkeiten kommt (weil die Leute Rechtschreibfehler machen, unsauber einzelne Assoziationen trennen, etc.).

Im Folgenden werden zuerst die Dateien und das Format beschrieben und dann die Korrekturen, die eingearbeitet werden sollen.

Danke, dass du das übernimmst! ☺

1. Dateien

Es gibt zwei Ordner, *komposita* und *namen*, in denen 5 bzw. 2 einzelne csv-Dateien liegen.

Die komposita-Dateien haben folgendes Format:

id	target_pnc	target_name	raw-associations	correction
5dea808cce8d8d19f5424b21	Abschiebe-Kretschmann	Winfried Kretschmann	Migration, Grüne, BaWü	Migration, Grüne, BaWü
6154d1ad6039a7b481511037	Abschiebe-Kretschmann	Winfried Kretschmann	Grüne, Ministerpräsident, Schwabe	Grüne, Ministerpräsident, Schwabe

Die namen-Dateien sehen sehr ähnlich aus, hier fehlt jedoch das Personal Name Compound:

id	target_name	raw-associations	correction
5ba1606a7afca3000194ff8a	Alexander Gauland	AFD, Faschist, Nazi	AFD, Faschist, Nazi
631b1380697762820b818b6a	Alexander Gauland	Politiker, AFD	Politiker, AFD

Bitte trage deine Korrekturen **ausschließlich** in der Spalte „correction“ (hier grün hinterlegt) neben der Spalte „raw-associations“ ein.
Die Spalte „raw-associations“-Spalte kann ggf. zum Abgleich genutzt werden, bitte trage dort jedoch **keine** Korrekturen ein.

Du kannst die csv-Dateien gerne in Excel oder Numbers bearbeiten, bitte exportiere die Daten am Ende jedoch wieder in csv-Dateien.

2. Korrekturen

Ziel aller Korrekturen ist, Assoziationen zu erhalten, die maschinell einles- und verwertbar sind, miteinander verglichen werden können, und sich tatsächlich auf den Namen oder das Kompositum beziehen.

Jede Assoziationen-Zelle soll als String eingelesen und am Komma getrennt werden, um die einzelnen Assoziationen zu erhalten. Wir möchten dann z.B. auswerten, ob viele Menschen zu einer Person ähnliche Assoziationen haben, etc. Dabei werden wir natürlich auch NLP-Methoden wie Stemming oder semantische Ähnlichkeits-Bestimmungen einsetzen, aber die Grundlage für unser Weiterarbeiten ist, dass die Assoziationen orthografisch so korrekt wie möglich sind, sich tatsächlich auf das Target (Name/Kompositum) beziehen und am Komma trennbar sind.

Abgesehen von einigen wenigen Änderungen meinerseits, sind die Assoziationen wie sie von den Annotator:innen kamen.

- Wenn Assoziationen nicht mit einem Komma, sondern einem Leerzeichen, Semikolon, Punkten, etc. getrennt wurden, bitte Kommata einfügen. Manche Menschen haben auch eine Aufzählung (1. …, 2. …) gemacht. Diese bitte ebenfalls auflösen und Kommas einfügen.

  Beispiele:
  -- „CDU Bundestag neoliberal“ → „CDU, Bundestag, neoliberal“
  -- „Medienkampagne, Altbundeskanzler. Was ist erlaubt, wenn es der Masse nicht gefällt? → Medienkampagne, Altbundeskanzler, Was ist erlaubt, wenn es der Masse nicht gefällt?

- Einige wenige Menschen haben sich viel Mühe gegeben und lange, korrekt mit Komma getrennte Sätze geschrieben. Bitte diese Kommas (also in ganzen Sätzen oder Phrasen mit Kommas, die jedoch als eine zusammenhängende Assoziation angegeben wurden) rausnehmen, damit diese Sätze / Phrasen nicht gesplittet werden. Sollte die Aufsplittung sinnvoll sein, können Kommas dringelassen werden.

  Beispiele:
  "Ich denke es ist eine Anspielung auf die Tatsache, das er im Rollstuhl saß. Was natürlich im Gedächtnis bleibt, ist seine Verstrickung in die CDU-Spendenaffäre." → "Ich denke es ist eine Anspielung auf die Tatsache dass er im Rollstuhl saß Was natürlich im Gedächtnis bleibt ist seine Verstrickung in die CDU-Spendenaffäre"
  → "Sie hat ihren Doktortitel erkuaft und von jemand anderem die Doktorarbeit schreiben lassen, bzw sie abgeschrieben" → "Sie hat ihren Doktortitel erkauft und von jemand anderem die Doktorarbeit schreiben lassen beziehungsweise sie abgeschrieben"

- Ausgeschriebene Trennzeichen wie „und“ und „entweder … oder“ bitte in Komma umwandeln, wenn im Sinne eines solchen benutzt

Beispiele:
-- „entweder ein 2. Vorname oder die Sicherheitsfirma von der er sich mal hat anwerben lassen“ → "ein 2. Vorname", "die Sicherheitsfirma von der er sich mal hat anwerben lassen"
-- "Deutsche Meisterschaft und DFB-Pokal" → "Deutsche Meisterschaft", "DFB-Pokal"
-- "legendär und lässig" →"legendär", "lässig"
-- "Schwarz und Stolz“ → „Schwarz“, stolz“
-- „Invasion in Afghanistan und Irak --> „Invasion in Afghanistan“, „Invasion in Irak“

- Punkte am Ende einer Aufzählung löschen

  Beispiele:
  -- „CDU, Bundestag, neoliberal.“ → „CDU, Bundestag, neoliberal“

- Satzendzeichen (wenn als solche genutzt) löschen (.?!)

  Wenn Menschen ein Fragezeichen (?) nutzen, um anzugeben, dass sie nicht 100% sicher sind, dass ihre Assoziation korrekt ist, das ? bitte erst einmal drin lassen.

- Abkürzungen auflösen bzw. entfernen

  Beispiele:
  -- „Türk. Präsident“ → „Türkischer Präsident“
  -- "dt. Nationalmannschaft" → „deutsche Nationalmannschaft“
  oder entfernen:
  -- „ex. Vizekanzler --> Ex-Vizekanzler“
  oder an die korrekte Stelle bringen:
  -- Bundeskanzlerin (ehem.) --> „ehemalige Bundeskanzlerin“

- Rechtschreibfehler jeglicher Art korrigieren

  Beispiele:
  -- "zwilichter Charakter" → "zwielichtiger Charakter"
  -- "Wilkommenskultur" → "Willkommenskultur
  -- "Ex-Bundeskanlerin" → "Ex-Bundeskanzlerin
  -- „Sehr erfolgreich e Skifahrerin“ → „Sehr erfolgreiche Skifahrerin“
  -- „Er ist sehr verletzt in der Hochfinanz und in der Wirtschaft macht gerne Aktien siehe Blackrock“ → „Er ist sehr vernetzt in der Hochfinanz und in der Wirtschaft macht gerne Aktien siehe Blackrock“

- Umlaute und ss bitte in die Standard-Schreibweise umwandeln:

  Beispiele:
  „Fussballer“ → „Fußball“

„Waehrung“ → „Währung“

- „&“ bitte in „und“ umwandeln.

- Fehlende oder überflüssige Bindestriche bitte einfügen/entfernen:

  Beispiele:
  -- „Ex Bundeskanzler“ → „Ex-Bundeskanzler“
  -- „Formel-1“ → „Formel 1“

- Kommentare wie "(keine weiteren Assoziationen)", "(wo soll ich 3-5 Assoziationen hernehmen?!)" ohne Substitution entfernen. Wenn durch Entfernung eines Kommentars keine Assoziationen zurückbleiben, bitte SPITZNAMEN IN ASSOZIATION ALS UNBEKANNT ANGEGEBEN eintragen.

- Wenn Menschen angeben, dass sie die Person oder das Kompositum nicht kennen, die Assoziationen bitte entfernen und „SPITZNAMEN IN ASSOZIATION ALS UNBEKANNT ANGEGEBEN“ eintragen.

  Bitte auch umwandeln, wenn Menschen sagen, dass sie den Spitznamen“ nicht kennen und dann Assoziationen zum Namen angeben.

  Beispiele:
  -- kenne den Spitznamen: Ferkel-Merkel nicht
  -- Kein Bezug zum Spitznamen.
  -- Ich kenne Ruhrpott-Messi nicht

- Manche Namen haben viele zugeordnete Personal Compounds, z.B. Markus Söder. Manche Annotator:innen waren da nicht so glücklich damit und äußerten das in Phrasen wie „schon wieder Söder“. Bitte ohne Substitution entfernen.

- Wenn Menschen „Kenne ich nicht so gut“, o. Ä. schreiben und dann eine Assoziation angeben, „Kenne ich nicht so gut“ entfernen und die Assoziation übernehmen.

- Manche Menschen waren über die Komposita unglücklich und fanden diese „unpassend“ oder „respektlos“ oder äußerten, dass die Person diesen Spitznamen nicht verdient habe. Bitte diese Assoziationen entfernen, da sie sich nicht auf den Spitznamen an sich beziehen und „MEINUNG ZU SPITZNAMEN-VERGABE“ eintragen.

Sollten wir sinnvoll Korrekturen vergessen haben, bzw. dir Dinge auffallen, schreibe jederzeit gerne eine E-Mail, damit wir kurz überlegen können, ob das Sinn ergebt aufzunehmen und dokumentiere die zusätzlichen Änderungen bitte in einem README, dass du am Ende mitschickst. Vielen lieben Dank!