Guide til annotering af tekstdokumenter

Oversigt

Dette dokument indeholder retningslinjer for annotering af tekst materiale, således at det bliver anonymiseret mht. GDPR lovgivningen. Du får en række tekstdokumenter, hvori der optræder en eller flere personer, hvis identitet skal beskyttes. Målet med din annotering er at vurdere, hvilke dele af teksten, som skal maskeres, så teksten bliver anonymiseret. *Læs venligst instruktionerne nedenfor grundigt, inden du går i gang med at annotere.*

For hvert dokument består din opgave af 4 trin, som vil hjælpe dig igennem processen:

- **Trin 0: Gennemlæsning.** Læs hele dokumentet igennem.
- **Trin 1: Tekstenheder.** I trin 1 skal du annotere alle *tekstenheder* du kan finde, som svarer til en af 8 *semantiske kategorier* (kan ses i tabellen under trin 1).
- **Trin 2: Maskering.** For hver tekstenhed, som du har annoteret i trin 1, skal du nu angive hvorvidt den skal maskeres, såfremt den kan bruges til at direkte- eller indirekte identificere en person (tekstenhedens *identifikator-type*), samt hvorvidt tekstenheden falder under en af 5 *fortrolige kategorier* (kan ses i tabellerne under trin 2).
- **Trin 3: Sidste gennemgang.** Når du er færdig med din annotering skal du til sidst gennemgå dokumentet for at sikre, at du ikke har overset nogle tekstenheder. Såfremt du opdager fejl/mangler, skal du korrigere dem ved at gennemgå trin 1 og 2 igen.

Trin 1: Tekstenheder

I trin 1 skal du finde og markere tekstenheder. En tekstenhed er en strækning af et eller flere tegn og/eller ord, som tilsammen udgør en enhed, der udgør en information om et individ. Disse informationer passer typisk i en semantisk kategori. Listen over semantiske kategorier præsenteres i tabellen nedenfor. Hvis tekstenheden ikke passer ind i en af disse kategorier, skal MISC-kategorien bruges (se *Eksempler på semantiske kategorier*).

Semantisk kategori	Beskrivelse
PERSON	Navne på personer, herunder aliaser/kaldenavne, brugernavne og initialer
CODE	Omfatter ID-numre og koder, der identificerer noget, såsom CPR, telefonnummer, pasnummer, nummerplade osv.
LOC	Omfatter byer, områder, lande, adresser samt navngivne infrastrukturer som busstoppesteder, broer osv.
ORG	Omfatter navne på organisationer, virksomheder, offentlige institutioner, skoler, NGO'er osv.
DEM	Demografiske egenskaber som alder, etnicitet, jobtitler, uddannelse, fysiske beskrivelser.
DATETIME	Specifikke datoer, tidspunkter eller varigheder.
QUANTITY	Betydningsfulde mængder, som procentdele eller monetære værdier.
MISC	Alle andre oplysninger, der beskriver en person, men ikke falder ind under de andre kategorier.

For at gøre din opgave nemmere er dokumenterne blevet præ-annoteret - det vil sige, at nogle tekstenheder allerede er blevet markeret (dog er der *ikke* angivet identifikator-type).

OBS: ***Præ-annoteringerne er fejlbarlige, og er blot et udgangspunkt for din annotering - du skal aktivt rette i annoteringerne samt tilføje dine egne.***

Sådan gør du:

1. **Kontroller præ-annoteret tekstenheder.** Tekstenheder er markeret med en farve og kode, som viser hvilken semantisk kategori de tilhører. Klik på en tekstenhed for at ændre den. Kontroller, om hver præ-annoteret tekstenhed er annoteret korrekt, eller skal fjernes/redigeres, eksempelvis fordi den semantiske kategori er forkert eller ordstrækningen skal ændres.
2. **Annoter nye tekstenheder.** Annoter herefter nye tekstenheder, der ikke er blevet opdaget af det automatiske værktøj. En tekstenhed skal passe i en af de 8 semantiske kategorier. Vælg først en kategori ved at klikke på den i venstre side af displayet, eller brug tal-knapperne på tastaturet til hurtigt at vælge kategori. Marker herefter tekstenheden med musen, og tekstenheden markeres med den farve og kode, som hører til kategorien.

Når du annoterer, skal du generelt markere den *mindste ordstrækning*, der angiver den pågældende tekstenhed. Dette betyder eksempelvis, at et mellemrum efter det sidste ord i ordstrækningen ikke skal inkluderes i tekstenheden.

I tilfælde af, at to eller flere tekstenheder refererer til det samme underliggende tekstenhed (f.eks. "Hans Jørgensen" og "Hr. Jørgensen"), skal du oprette en relation mellem tekstenhederne (se mere i afsnittet *Relationer* under *Eksempler på kategorier*).

OBS: I trin 1 behøver du ikke bekymre dig om hvorvidt tekstenhederne kan bruges til re-identifikation af personer (det kommer i Trin 2), du skal blot annotere alle tekstenheder, der tilhører en af de semantiske kategorier.

Eksempler på semantiske kategorier

PERSON

For personnavne skal annoteringen inkludere titler og tiltaleformer, såsom Hr., Dr., osv., da disse kan bidrage til at identificere en person.

- Hr. Gestur Jónsson$_{(PERSON)}$ og Hr. Ragnar Halldór Hall $_{(PERSON)}$
- Hr. og Fru Jørgensen$_{(PERSON)}$

Eksempler på, hvad der betragtes som navne:

- **Navne:** F.eks. 'Hans Jørgensen, 'Jørgensen', 'Hans'
- **Initialer:** F.eks. 'H.H.'
- **Stavefejl:** F.eks. 'Hans Jørnsen'
- **Alle ortografiske variationer:** F.eks. 'hans jørgensen', 'Hans JØRGENSEN'

CODE

Omfatter ID-numre, koder og andre talrækker, der identificerer noget, såsom CPR-nummer, telefonnummer, pasnummer, nummerplade, rapportnumre osv.

- Angående sagsnummer (nr. 42552/98)$_{(CODE)}$

LOC

Omfatter byer, områder, kommuner, adresser samt andre geografiske steder, bygninger og faciliteter. Andre eksempler er lufthavne, kirker, restauranter, hoteller, turistattraktioner, hospitaler, butikker, adresser, veje, have, fjorde, bjerge og parker.

- Reykjavik$_{(LOC)}$
- Øvregaten 2a, 5003 Bergen$_{(LOC)}$

Inkludér tal, når de er en del af navnet:

- Pilestredet 48$_{(LOC)}$
- Rema 1000$_{(ORG)}$

Annoter altid hele ordstrækningen, selv hvis flere ord indgår i samme tekstenhed:

- Høgskolen i Oslo og Akershus$_{(ORG)}$

Hvis der er tale om separate tekstenheder forbundet med en konjunktion (f.eks. "og"), skal de annoteres separat:

- Gågaderne i Viborg og Silkeborg$_{(LOC)}$

ORG

Omfatter enhver navngiven samling af mennesker, virksomheder, institutioner, organisationer, universiteter, hospitaler, kirker, sportshold, fagforeninger, politiske partier osv.

Firmaangivelser som AS, Co. og Ltd. skal inkluderes som en del af navnet.

- A.P. Møller - Mærsk A/S$_{(ORG)}$

Oversættelser og akronymer inkluderes i markeringen, f.eks:

- Aarhus Universitet (AU)$_{(ORG)}$

Bestemte eller ubestemte artikler medtages typisk ikke i markeringen, medmindre de eksplicit er en del af tekstenheden.

- Det Danske Sprog- og Litteraturselskab$_{(ORG)}$
- Sidste års vinder af Den Store Bagedyst$_{(MISC)}$

Hvis en tekstenhed kan tilhører flere kategorier, skal du vælge den semantiske kategori, der bedst beskriver enheden ud fra konteksten:

- Sverige$_{(ORG)}$ har opkøbt en strategisk Bitcoin reserve
- Sverige$_{(LOC)}$ ligger øst for Norge$_{(LOC)}$

DEM

Disse er demografiske markører, og omfatter både fysiske, kulturelle og erhvervsmæssige/uddannelsesmæssige attributter, såsom fysiske beskrivelser, diagnoser, modersmål, etnicitet, jobtitler, alder osv.

- 40 år$_{(DEM)}$ gammel

- ansøgeren er journalist(DEM)
- en gruppe venstreorienterede(DEM) ekstremister
- diagnosticeret med brystkræft(DEM)
- en svensk(DEM) fysiker

Pronominer (han, hun) skal ikke medtages i tekstenhederne.

DATETIME

Præpositioner (f.eks. på, ved) skal ikke inkluderes i markeringen.

- Mandag, 3. oktober 2018(DATETIME)
- kl. 9:48(DATETIME)
- født i 1947(DATETIME)

Separate tekstenheder forbundet med "og" skal annoteres separat:

- 10. marts(DATETIME) og 12. marts(DATETIME)

Såfremt tekstenhederne ikke kan separeres skal de annoteres samlet:

- 10. og 12. marts(DATETIME)

QUANTITY

Meningsfulde mængder, såsom valutaer. Enheden (eks. valutaen) skal inkluderes i ordstrækningen.

- 37.5 millioner kroner(QUANTITY)
- 375 euro(QUANTITY)
- 4267 SEK(QUANTITY)
- 1000 kilo(QUANTITY)

MISC

Andre tekstenheder der kan anses som enheder, såsom varemærker, produkter, kunstværker, begivenheder osv. Alle kunstigt fremstillede ting betragtes som produkter. Dette kan også inkludere mere abstrakte tekstenheder såsom taler, radioprogrammer, programmeringssprog, kontrakter, lovgivning og teorier (såfremt de er navngivet).

Brands er produkter (or derfor MISC), når de henviser til et produkt eller en produktlinje, men organisationer (ORG), når de henviser til handlende eller producerende instanser.

- Lego(MISC) er et af verdens mest populære legetøjsmærker
- Har du set Mona Lisa(MISC)?
- Jeg glæder mig til Roskilde Festival(MISC)!

Relationer

Hvis nogle tekstenheder refererer til den samme underliggende tekstenhed gennem forskellige formuleringer, skal du oprette en relation mellem tekstenhederne:

1. Klik på den seneste omtale (f.eks. "John Smith"), og klik herefter på *Create relation between regions*-knappen, som er vist med et link-ikon i højre side af interfacet.
2. Find forekomsten af den anden omtale (f.eks. "Smith, John") og klik på den. Du burde nu se en relation mellem de to tekstenheder, markeret med en pil.

OBS: Du skal kun oprette en relation, såfremt tekstenhederne indeholder samme mængde information. Eksempelvis er "Jørgen Hansen" og "Jørgen" ikke lige informative, og afhængigt af konteksten kan førstnævnte ses som en direkte identifikator og sidstnævnte en indirekte identifikator.

Trin 2: Maskering

I trin 2 skal du for hver tekstenhed markeret i trin 1 angive, om denne skal **maskeres**, såfremt den kan bruges til **direkte- eller indirekte at identificere personer**, som indgår i teksten.

Mere præcist skal du for hver tekstenhed annoteret i trin 1 angive **identifikator-type**, som består af følgende typer:

Identifikator-type	Beskrivelse
DIREKTE	Såfremt tekstenheden udgør direkte og utvetydigt identificerende information, skal du sætte kryds ved DIREKTE.
KVASI	Såfremt tekstenheden udgør indirekte identificerende information, der i kombination med baggrundsviden og andre indirekte identifikatorer kan identificere en person, skal du sætte kryds ved KVASI. Indirekte identfikatorer kaldes også for kvasi-identifikatorer.

Når du er gået til trin 2 vil du kunne se to nye kategorier i højre side: **DIREKTE** og **KVASI**. Såfremt tekstenheden er en direkte- eller kvasi identifikator, skal du ændre kategorien ved først at klikke på tekstenheden, så den lyser op, og derefter vælge den nye kategori. Tekstenheden vil nu blive maskeret - direkte identifikatorer er sorte, og kvasi identifikatorer er grå.

Eksempel:

Dette er en direkte identifikator$_{(PERSON)}$, og dette er en kvasi identifikator$_{(MISC)}$.

Efter at have valgt identifikator-type:

Dette er en [redacted], og dette er en [redacted].

Hvis tekstenheden ikke er en direkte- eller kvasi identifikator, skal du ikke ændre på tekstenheden.

Overvej nøje, om en indirekte identifikator bør maskeres!

I mange tilfælde er det nemlig ikke nødvendigt at maskere alle indirekte identifikatorer for at beskytte individets identitet. Hvis der er mange potentielle indirekte identifikatorer, skal du maskere så få som muligt imens du sikre dig, at de resterende tekstenheder ikke udgør nok information til at identificere individet. Der er ofte flere kombinationer af indirekte identifikatorer som kan maskeres, og du skal derfor vurdere, hvilke du kan lade forblive synlige i teksten.

Hvad er en person?

En person er her defineret som en naturlig person, hvilket omfatter alle **levende, menneskelige individer**. Du skal være opmærksom på, at dokumentet først er anonymiseret idet identiteten af alle personer, som optræder i teksten, er beskyttet.

Udover identifikator-type skal du angive, hvorvidt informationen i tekstenheden tilhører en **fortrolig kategori**. Du kan angive hvorvidt informationen falder i en af de fortrolige kategorier ved af sætte kryds ud fra kategorien til venstre i interfacet. Informationen er fortrolig såfremt hvis den falder under en af følgende kategorier:

Fortrolig kategori	Beskrivelse
BELIEF	Religiøse eller filosofiske overbevisninger
POLITICS	Politiske holdninger eller tilhørsforhold
SEX	Sexuel orientering elle andre oplysninger om sexliv
ETHNIC	Etnisk oprindelse
HEALTH	Sundheds-, genetisk- og biometrisk data, herunder sensitive sundhedsmæssige oplysninger som misbrug o. lign.

Hvis informationen ikke er fortrolig, skal du ikke sætte et kryds.

Du kan maskere tekstenheder, som du ikke har opdaget i trin 1.

Såfremt du opdager en tekstenhed, som du ikke har markeret i Trin 1, og som er en direkte- eller indirekte identifikator, kan du markere den i Trin 2 for at maskere den med det samme.

Eksempler på identifikatorer

Direkte identifikatorer

Information, som direkte og utvetydigt kan identificere en person. Typiske eksempler er personnavne (herunder kælenavne/aliasser og brugernavne), CPR-numre, telefonnumre, e-mailadresser, adresser, kontooplysninger mm.

Et personnavn kan enten være en direkte identifikator (hvis det er det fulde navn på en person) eller en indirekte identifikator, afhængigt af konteksten. Eksempelvis er nogle navne så udbredte, at de ikke leder til direkte identifikation af en person. Ligeledes kan diverse koder være enten direkte- eller indirekte identifikatorer, alt efter om de entydigt refererer til den person, der skal beskyttes, eller ej.

Indirekte (kvasi) identifikatorer

Information, der i sig selv ikke kan bruges til at identificerer en person, men som i kombination med andre indirekte identifikatorer og/eller baggrundsviden kan lede til identificering af en person. Indirekte identifikatorer kan eksempelvis være demografiske oplysninger (“en 72-årig mand”) eller angivelser af tid og/eller sted (“den 6. februar i Sevilla”). En kombination af fødselsdato, køn og erhverv vil typisk gøre det muligt at finde frem til en persons identitet.

For at indirekte identifikatorer kan lede til identificering af en person, skal disse ses i sammenhæng med "offentligt tilgængelig viden"; oplysninger, som man med rimelighed kan forvente, at en ekstern person enten allerede ved om individet, eller vil kunne finde ud af. Med andre ord bør du stille dig selv følgende spørgsmål: Hvis jeg ville finde ud af, hvem personen i dokumentet er, ville jeg så være i stand til at kombinere disse informationer med offentligt tilgængelige kilder (såsom nyhedsartikler, sociale medier, databaser, osv.)? Og er disse oplysninger da nok

til at re-identificere individet med ingen eller lav tvetydighed? *I langt de fleste tilfælde behøver du ikke at undersøge disse kilder — din umiddelbare intuition er nok*. Hvis du vurderer, at kombinationen af indirekte identifikatorer samt offentligt tilgængelig viden ikke er tilstrækkelig for at re-identificere individet med ingen eller lav usikkerhed, kan tekstenhederne da ikke anses for at være indirekte identifikatorer.

Hvad er offentligt tilgængelig viden? I praksis omfatter det alt, der kan findes ved at søge på internettet. Dog er en del af de dokumenter, du skal annotere, hentet fra internettet, og i disse tilfælde er dokumentet selv ikke omfattet i definition af offentligt tilgængelig viden; her skal man forestille sig, at dokumentet ikke er tilgængeligt på internettet.

Som tommelfingerregel bør uforanderlige personlige attributter (såsom fødselsdato), som kan kendes eller forefindes ved eksterne kilder/databaser, betragtes som indirekte identifikatorer. Andre personlige attributter kan betragtes som indirekte identifikatorer afhængigt af sandsynligheden for, at disse oplysninger kan kendes eller forefindes ved eksterne kilder/databaser. Eksempelvis kan nuværende bopæl eller dato for en hospitalsindlæggelse indirekte identificere en person, men antallet af gange man har handlet i supermarkedet på en uge vil sjældent kunne anses som en indirekte identifikator. Kun meget generelle attributter, som kendetegner et stort antal personer (f.eks. fødeland), ignoreres, da de ofte ikke markant øger muligheden for re-identifikation af en person. Dog er konteksten yderst vigtig, herunder hvilke andre indirekte identifikatorer der findes i dokumentet. Jo flere, der findes, og jo mere konkrete oplysninger de giver, desto større er sandsynligheden for, at de tilsammen kan muliggøre re-identifikation af en person.

Trin 3: Sidste gennemgang

Når du er færdig med trin 1 og 2 skal du til sidst gennemgå dokumentet for at sikre, at du ikke har overset nogle tekstenheder. Ved dette trin skal du trykke på CTRL + H i Label Studio for at skjule de tekstenheder, som du ikke har vurderet til at være direkte- eller indirekte identifikatorer.

Læs dokumentet igennem igen - ville du nu være i stand til at re-identificere en eller flere personer? Hvis ja, skal du gennemgå trin 1 og 2 igen, indtil dette ikke længere er tilfældet.

Når du er færdig skal du gemme dokumentet ved at klikke på Submit-knappen. Du kan nu fortsætte til næste dokument.

Eksporter annoteringer

For at eksportere dine annoteringer skal du klikke på Export-knappen i højre hjørne når projektet er åbent i Label Studio. Gem som JSON-format.