Textsegmentering

Inför laboration 1

729G49 Språk och Datorer

(baserat på Marco Kuhlmanns bilder från 2020)

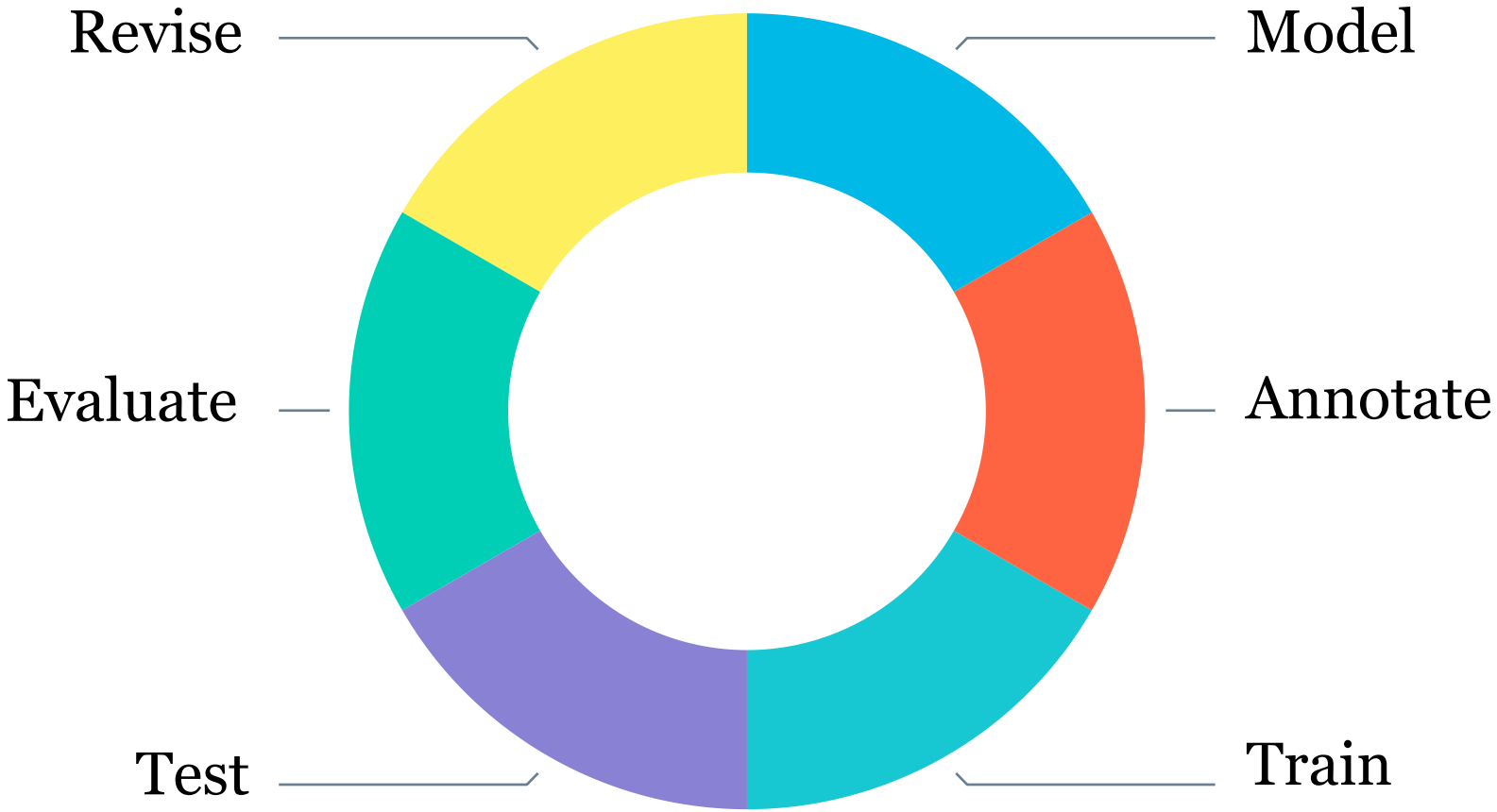
LINKÖPINGS UNIVERSITET

Denna föreläsning

- Att utveckla korpusar
- Från webbsida till segmenterad text
- Reguljära uttryck

Att utveckla korpusar

Utvecklingscykeln



Källa: Pustejovsky och Stubbs (2013)

Sentimentanalys

The gorgeously elaborate continuation of “The Lord of the Rings” trilogy is so huge that a column of words cannot adequately describe co-writer/director Peter Jackson’s expanded vision of J.R.R. Tolkien’s Middle-earth

… is a sour little movie at its core; an exploration of the emptiness that underlay the relentless gaiety of the 1920’s, as if to stop would hasten the economic and global political turmoil that was to come.

Hitta namngivna entiteter (named entities)

The gorgeously elaborate continuation of "The Lord of the Rings" trilogy is so huge that a column of words cannot adequately describe co-writer/director Peter Jackson's expanded vision of J.R.R. Tolkien's Middle-earth

… is a sour little movie at its core; an exploration of the emptiness that underlay the relentless gaiety of the 1920's, as if to stop would hasten the economic and global political turmoil that was to come.

Filmtitel Person Plats Tidsepok

Hitta namngivna entiteter (named entities)

The gorgeously elaborate continuation of "The Lord of the Rings" trilogy is so huge that a column of words cannot adequately describe co-writer/director Peter Jackson's expanded vision of J.R.R. Tolkien's Middle-earth

… is a sour little movie at its core; an exploration of the emptiness that underlay the relentless gaiety of the 1920's, as if to stop would hasten the economic and global political turmoil that was to come.

Filmtitel Person Plats Tidsepok

Hitta namngivna entiteter (named entities)

The gorgeously elaborate continuation of "The Lord of the Rings" trilogy is so huge that a column of words cannot adequately describe co-writer/director Peter Jackson's expanded vision of J.R.R. Tolkien's Middle-earth

… is a sour little movie at its core; an exploration of the emptiness that underlay the relentless gaiety of the 1920's, as if to stop would hasten the economic and global political turmoil that was to come.

Filmtitel Person Plats Tidsepok

Hitta namngivna entiteter (named entities)

The gorgeously elaborate continuation of "The Lord of the Rings" trilogy is so huge that a column of words cannot adequately describe co-writer/director Peter Jackson's expanded vision of J.R.R. Tolkien's Middle-earth

… is a sour little movie at its core; an exploration of the emptiness that underlay the relentless gaiety of the 1920's, as if to stop would hasten the economic and global political turmoil that was to come.

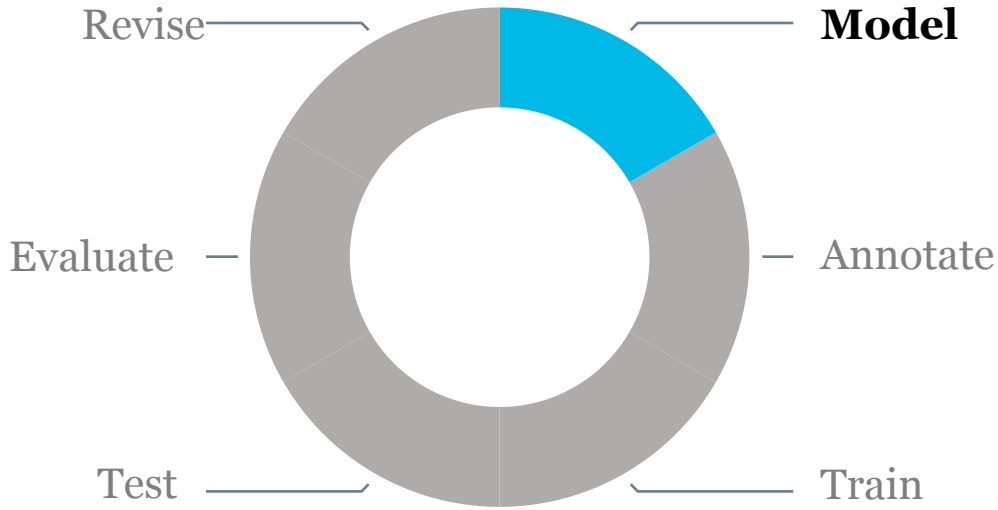
Filmtitel Person Plats Tidsepok

Steg 1: modellera

Modellen beskriver avsikten med korpusen i abstrakta termer

Sentimentanalys: Varje dokument tillhör en klass. Den kan uttrycka antingen en positiv eller en negativ attityd gentemot filmen som recenseras.

Hitta namngivna entiteter: Ordsekvenser i dokumentet kan beteckna namngivna entiteter, så som personer, platser eller tidsepoker

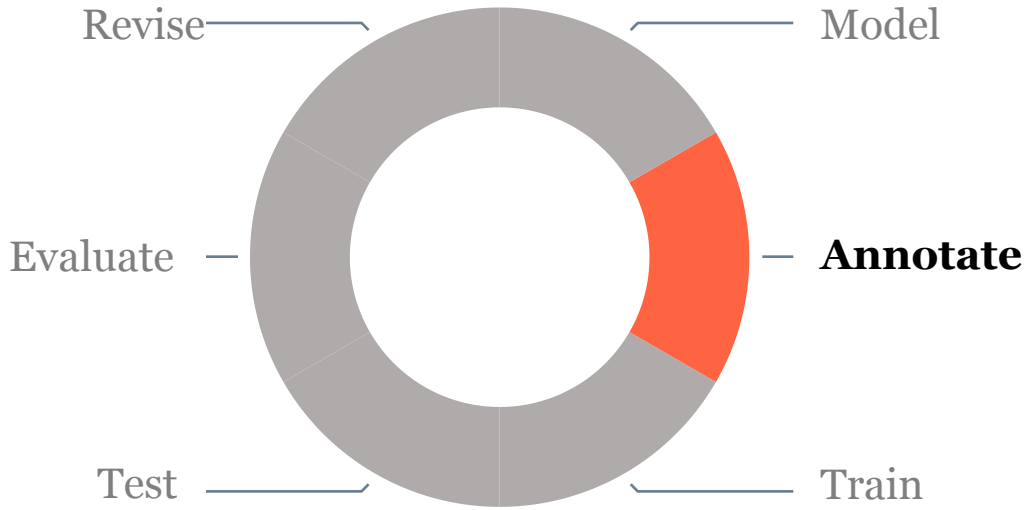


Steg 2: annotera

Datamängden annoteras utifrån annoteringsriktlinjer

Sentimentanalys: Annotera den övergripande attityden gentemot filmen som uttrycks i texten

Hitta namngivna entiteter: Annotera endast platser som finns i verkligheten, inte fiktiva platser så som *Middle-Earth*

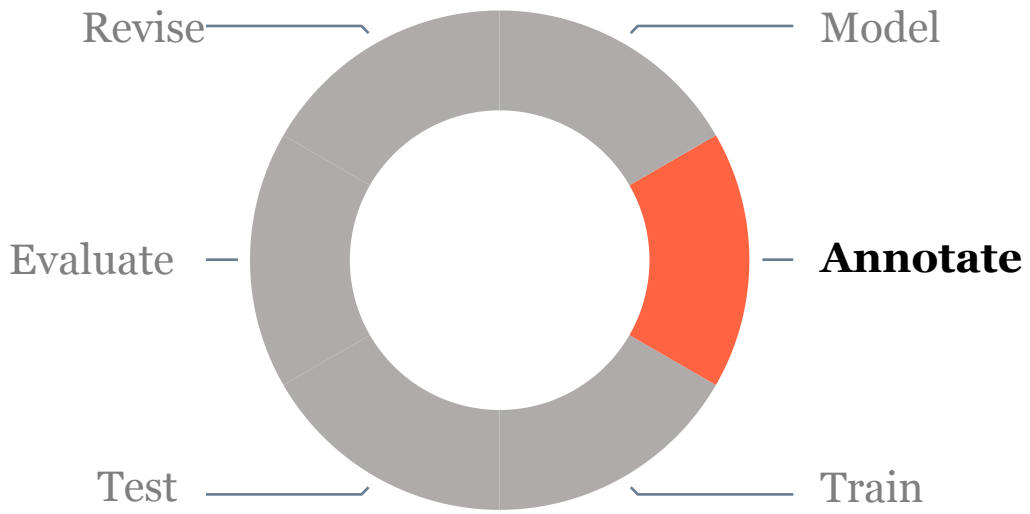


Steg 2: annotera

Textmaterialet annoteras av en eller flera personer. Dessa försöker följa riktlinjerna så noggrant som möjligt.

Annoteringsriktlinjerna diskuteras och kan behöva anpassas under annoteringsarbetet.

När textmaterialet har annoterats skapas en guldstandard genom att man jämför och diskuterar annoteringarna.



Steg 3 och 4: träna och testa

Guldstandardmängden delas in i två disjunkta delmängder

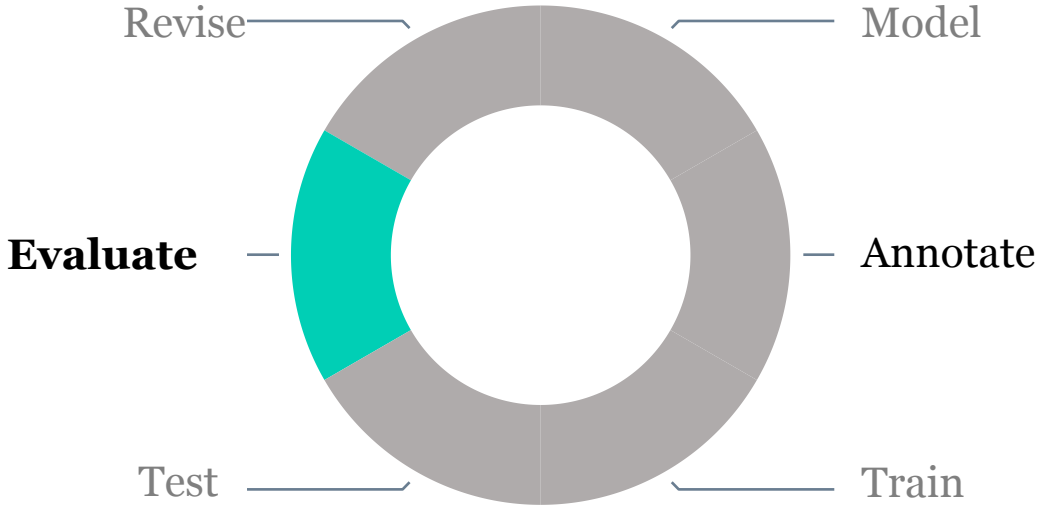
En **träningsmängd** som används för att träna upp ett system mha maskininlärning

En **testmängd** som används för att utvärdera det färdiga systemets förmåga att generalisera utanför träningsmängden

Steg 5: utvärdera

Hur väl presterar systemet?

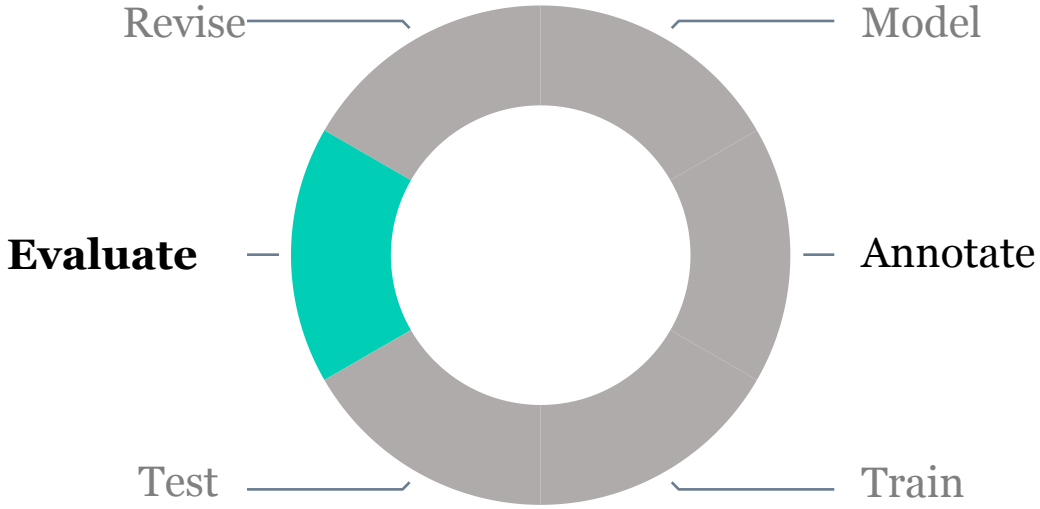
	Systemet: ”person”	**Systemet: ”inte person”**
Guldstandard: ”person”	sanna positiva	falska negativa
Guldstandard: ”inte person”	falska positiva	sanna negativa



Steg 5: utvärdera

Hur väl presterar systemet?

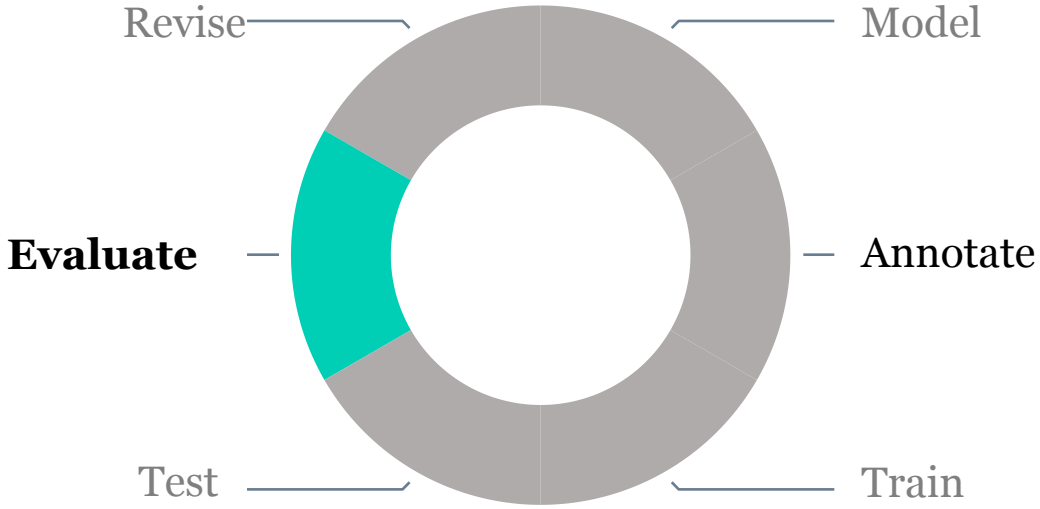
	Systemet: ”person”	**Systemet: ”inte person”**
Guldstandard: ”person”	sanna positiva	falska negativa
Guldstandard: ”inte person”	falska positiva	sanna negativa



Steg 5: utvärdera

$$\text{precision} = \frac{\text{sanna positiva}}{\text{sanna positiva} + \text{falska positiva}}$$

$$\text{täckning} = \frac{\text{sanna positiva}}{\text{sanna positiva} + \text{falska negativa}}$$



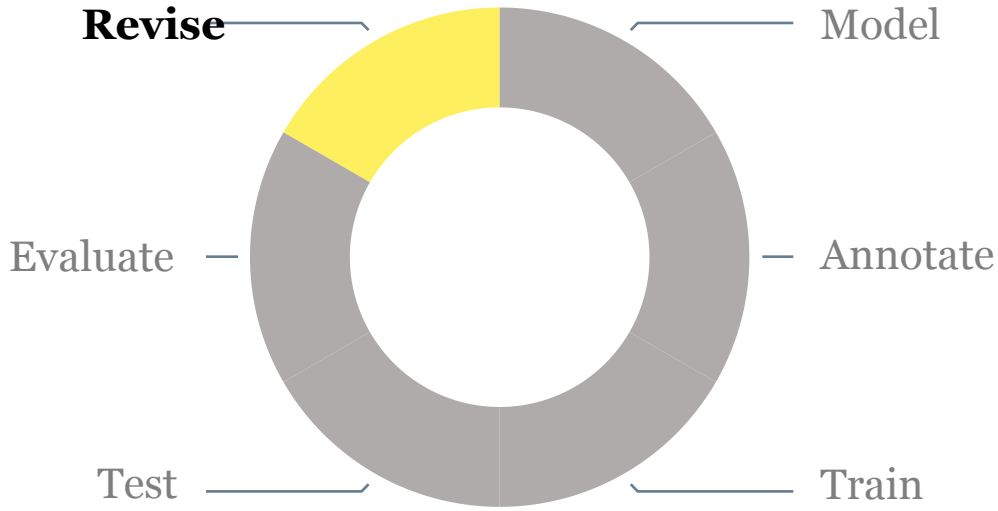
Steg 6: revidera

Som komplement till den kvantitativa utvärderingen bör man göra en kvalitativ utvärdering i form av en felanalys.

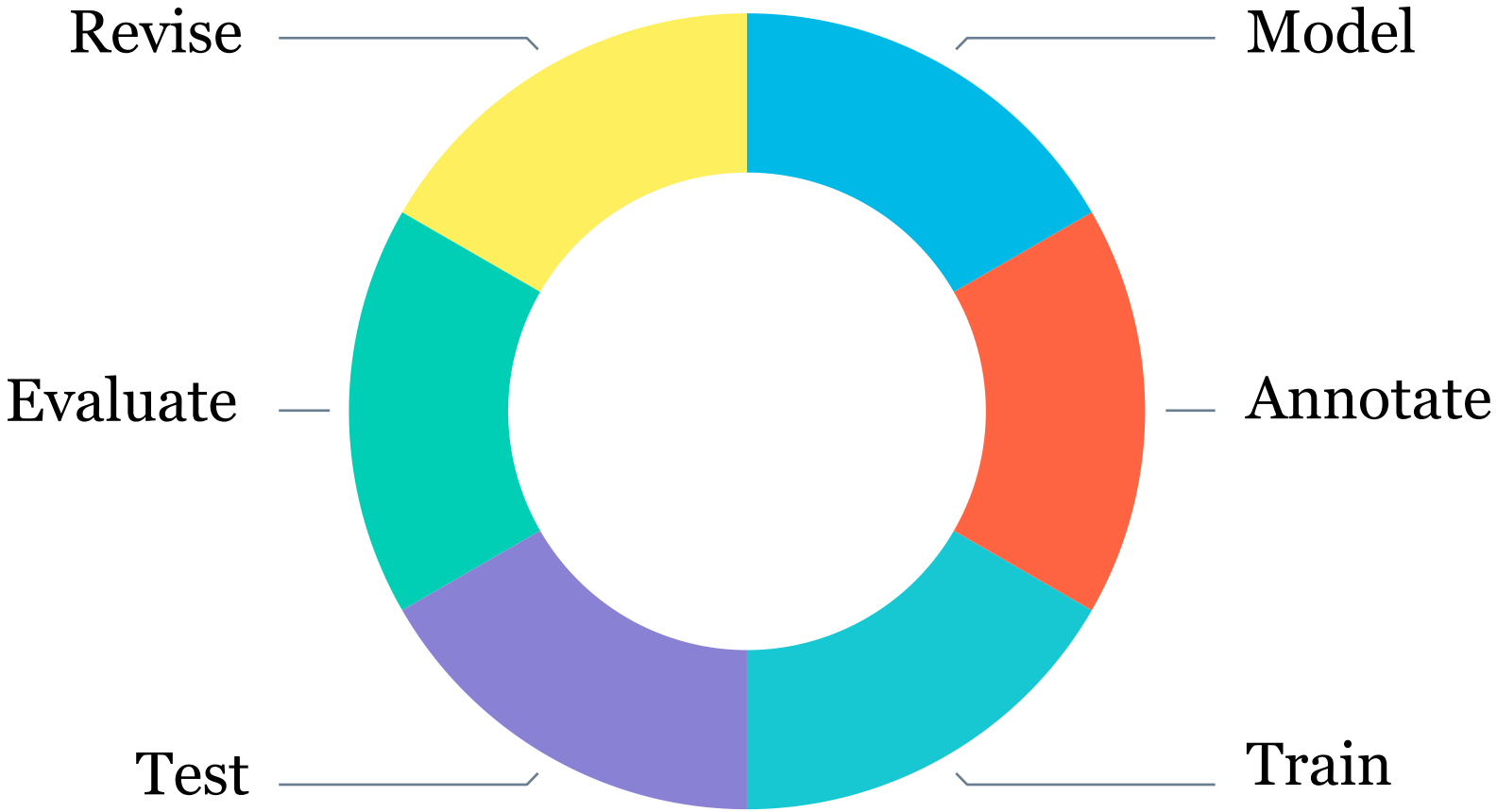
Exempel: vilka entiteter förväxlas oftast?

Denna felanalys kan leda till en ny modell, nya annoteringsriktlinjer och att annoteringen utvidgas

Exempel: Lägga till ordklasser



Utvecklingscykeln



Källa: Pustejovsky och Stubbs (2013)

Denna föreläsning

- Att utveckla korpusar
- Från webbsida till segmenterad text
- Reguljära uttryck

Från webbsida till segmenterad text

Från råmaterial till annoterad text

Aktivitet	Beskrivning
Urval	Välj ut de texter som ska vara med i korpusen
Insamling	Samla in texterna, t.ex. genom att ”spindla” webben
Avformatering	Ta bort strukturell uppmärkning, t.ex. XML
Segmentering	Dela upp textmaterialet i relevanta enheter
Annotering	Lägg till relevant information, t.ex. ordklasser

Exempel

https://sv.wikipedia.org/wiki/Kapybara

Textsegmentering

- **Textsegmentering** är uppgiften att dela upp en text i lingvistiskt meningsfulla enheter, t. ex. ord, meningar, stycken

- När enheterna som segmenteras är ord eller ordliknande enheter så kallas det för **tokenisering**

Tokenisering

Rå text

Den når en kroppslängd upp till 130 centimeter och en vikt upp till 61 kilogram. Arten förekommer i stora delar av Sydamerikas slättland samt i angränsande regioner av Centralamerika. Kapybaran jagas för köttets och hudens skull men den räknas inte till de hotade arterna. Artens närmaste släkting är klippmarsvinet och ibland listas de tillsammans som underfamilj till familjen marsvin (Caviidae).

Tokeniserad text

Den når en kroppslängd upp till 130 centimeter och en vikt upp till 61 kilogram . Arten förekommer i stora delar av Sydamerikas slättland samt i angränsande regioner av Centralamerika . Kapybaran jagas för köttets och hudens skull men den räknas inte till de hotade arterna . Artens närmaste släkting är klippmarsvinet och ibland listas de tillsammans som underfamilj till familjen marsvin (Caviidae) .

Möjliga fel vid tokenisering

Undersegmentering

Den automatiska tokeniseringen missar att segmentera en teckensekvens som enligt guldstandard ska segmenteras

Översegmentering

Den automatiska tokeniseringen delar på en teckensekvens som enligt guldstandarden inte ska segmenteras

bl. a.
t.ex.
New York
Anna-Lena
mat- och sovklocka

Meningssegmentering

- I vissa sammanhang vill vi inte bara identifiera ord utan även större enheter så som meningar eller stycken
- Meningssegmentering är uppgiften att dela upp en text i meningar
- Meningssegmentering är svårare än att dela upp en text efter en punkt eller något annat skiljetecken

  We visited the U.S. After that, we visited Canada.

Denna föreläsning

- Att utveckla korpusar
- Från webbsida till segmenterad text
- Reguljära uttryck

Reguljära uttryck

Tack för idag!

LINKÖPINGS UNIVERSITET