**Fundamentals of Text Analysis with the Language Service**

**Förstå Textanalys och NLP**

Textanalys är en metod för att extrahera information och mönster från text och används i många tillämpningar, från sentiment analysis till machine translation. Azure AI Language erbjuder verktyg för att hantera textanalys, men innan vi dyker in i dess funktioner är det bra att förstå några grundläggande koncept inom Natural Language Processing (NLP).

**Tokenization – Dela upp texten**

För att analysera en text börjar man ofta med att dela upp den i mindre enheter, kallade **tokens**. Vanligtvis motsvarar en token ett ord, men det kan också vara en del av ett ord eller en kombination av ord och skiljetecken.

Exempel: Meningen **"we choose to go to the moon"** kan delas upp i tokens som:

1. we
2. choose
3. to
4. go
5. the
6. moon

Man kan även förbättra tokenization genom att:

* **Text normalization** – ta bort skiljetecken och göra allt gemener.
* **Stop word removal** – filtrera bort vanliga men innehållslösa ord som "och", "eller", "det".
* **n-grams** – analysera ordkombinationer, t.ex. "New York" eller "machine learning".
* **Stemming** – förenkla ord till sin grundform, t.ex. att "running" och "ran" tolkas som "run".

**Frequency analysis – Hur ofta förekommer orden?**

Efter att ha brutit ner texten kan man räkna hur ofta varje token förekommer. Detta kan hjälpa till att förstå vad texten handlar om. Om man analyserar ett tal om månlandningen och ser att ord som "moon", "space" och "go" är vanligast, kan man dra slutsatsen att det handlar om rymdfärder.

För att förbättra analysen kan man använda **TF-IDF (Term Frequency - Inverse Document Frequency)**. Det jämför hur ofta ett ord förekommer i en specifik text jämfört med hur vanligt det är i andra texter. På så sätt kan vi identifiera vilka ord som är mest betydelsefulla i just den texten.

**Machine learning for text classification – Kategorisera text automatiskt**

Genom att använda machine learning kan vi automatiskt kategorisera text. Ett vanligt exempel är **sentiment analysis**, där vi försöker avgöra om en text är positiv eller negativ.

Exempel på restaurangrecensioner som kan klassificeras:

* *"The food and service were both great!"* → **1 (positiv)**
* *"Slow service and substandard food."* → **0 (negativ)**

Genom att träna en modell på många sådana exempel lär den sig känna igen vilka tokens som signalerar positiv eller negativ känsla.

**Semantic language models – Förstå meningen bakom orden**

Mer avancerade modeller använder **embeddings**, där varje token representeras av en flerdimensionell vektor. Tanken är att ord som liknar varandra placeras nära varandra i en matematisk rymd. Till exempel kommer "dog" och "cat" att ha liknande vektorer, medan "skateboard" hamnar längre bort.

Dessa language models kan användas för en rad olika NLP-uppgifter, exempelvis:

* **Text analysis** – hitta nyckelord och identifiera namn i en text.
* **Sentiment analysis** – avgöra känslan i en text.
* **Machine translation** – översätta text mellan språk.
* **Summarization** – extrahera de viktigaste delarna ur en lång text.
* **Conversational AI** – bygga chattbotar och digitala assistenter.

Azure AI Language erbjuder verktyg för dessa uppgifter och mer.

Här är en sammanfattning där alla tekniska termer behåller sitt engelska namn medan förklaringen är på svenska:

**Azure AI Language – Avancerad textanalys**

Azure AI Language är en del av Azure AI Services och används för att utföra avancerad **Natural Language Processing (NLP)** på ostrukturerad text. Verktyget innehåller flera funktioner för textanalys, såsom **Named Entity Recognition, Sentiment Analysis, Language Detection, Key Phrase Extraction** och **Summarization**.

**Entity Recognition och Entity Linking**

Med **Entity Recognition** kan Azure AI Language identifiera specifika enheter i texten, som personer, platser och organisationer. Systemet kan även tolka kvantiteter som siffror, procent, mått och valutor samt tidpunkter som datum, tider och tidsintervall.

Exempel på entiteter som kan identifieras:

* **Person:** "Bill Gates", "John"
* **Location:** "Paris", "New York"
* **Organization:** "Microsoft"
* **DateTime:** "May 2nd, 2017", "8am"
* **Quantity:** "25%", "10 miles", "45 degrees"
* **Email & URL:** "[support@microsoft.com](mailto:support@microsoft.com)", "[https://www.bing.com](https://www.bing.com/)"

För att undvika förväxlingar används **Entity Linking**, som kopplar igenkända entiteter till Wikipedia-artiklar för tydligare sammanhang.

Exempel:  
Text: *"I ate at the restaurant in Seattle last week."*  
Identifierad entitet: **Seattle** (Location) → <https://en.wikipedia.org/wiki/Seattle>

**Language Detection**

Azure AI Language kan identifiera språk i text och returnerar:

* Språkets namn (t.ex. "English")
* **ISO 639-1**-kod (t.ex. "en" för engelska, "es" för spanska)
* En **confidence score** som visar hur säker modellen är på sin bedömning.

Exempel på språkidentifiering i kundrecensioner:

| **Review** | **Language Name** | **ISO 639-1 Code** | **Score** |
| --- | --- | --- | --- |
| "A fantastic place for lunch." | English | en | 1.0 |
| "Comida maravillosa y gran servicio." | Spanish | es | 1.0 |
| "The croque monsieur avec frites was terrific. Bon appetit!" | English | en | 0.9 |

Blandspråk (som engelska och franska i Review 3) kan påverka confidence score, men systemet fokuserar på det dominerande språket.

**Sentiment Analysis och Opinion Mining**

Med **Sentiment Analysis** kan Azure AI Language avgöra om en text är positiv, neutral eller negativ. Varje mening får en poäng mellan 0 och 1 för varje sentimentkategori.

Exempel på sentimentanalys för restaurangrecensioner:

**Review 1:** *"The food was amazing and the staff was very friendly!"*

* **Document sentiment:** Positive
* **Positive score:** 0.90
* **Neutral score:** 0.10
* **Negative score:** 0.00

**Review 2:** *"The service was slow and the food was awful."*

* **Document sentiment:** Negative
* **Positive score:** 0.00
* **Neutral score:** 0.00
* **Negative score:** 0.99

**Opinion Mining** går ett steg längre genom att identifiera vilka aspekter i texten som bidrar till känslan.

**Key Phrase Extraction**

Funktionen **Key Phrase Extraction** extraherar nyckelbegrepp ur en text för att snabbt sammanfatta huvudpunkterna.

Exempel:

Text: *"We had dinner here for a birthday celebration and had a fantastic experience. The ambiance was relaxed, the food was amazing, and service was terrific."*

Nyckelbegrepp som identifieras:

* **birthday celebration**
* **fantastic experience**
* **ambiance**
* **great food**
* **attentive service**

Detta gör det enklare att analysera stora mängder text snabbt och identifiera återkommande teman.

**Skapa en Azure AI Language Resource**

För att använda Azure AI Language i en applikation behöver du skapa en resurs i Azure. Det finns två alternativ:

1. **Language resource** – Om du endast vill använda Azure AI Language och hantera åtkomst och fakturering separat.
2. **Azure AI Services resource** – Om du vill kombinera Azure AI Language med andra AI-tjänster och hantera allt i samma resurs.

Med Azure AI Language kan du effektivisera textanalys genom att automatisera språkidentifiering, entitetsigenkänning, sentimentanalys och nyckelordsutvinning.