# TFIDF(tf-idf)

- tf = term frequency
- idf = inverse document frequency
- Perustuu samojen taajuuksien laskentaan
- Pyritään löytämään tekstistä merkityksellisiä sanoja(features) eli avainsanoja

- Ensiksi lasketaan termin esiintymiskerrat dokumentissa/sanojen kokonaismäärä( vertaa count vectorizer)
- Sitten lasketaan termin esiintyminen koko datassa/corpuksessa(training data)
- Jos sana ilmenee useassa dokumentiissa -> sana on yleissana jolla on kohtuu vähän merkitystä tekstin asiasisältöön
- Toisaalta jos sana esiintyy usein harvoissa teksteissä-> sanan merkitys tekstiin on korkea eli se on avainsana

- score = tf * log(idf)

* tf = tietyn sanan määrä yksittäisessä dokumentissa/sanojen kokonaismäärällä

* idf = dokumenttien kokonaismäärä / monessako dokumentissa sana esiintyy