Cheat Sheet til søgning i Mediestream

Udarbejdet af Max Odsbjerg Pedersen

Eksempel 1:

Poul Schlüter

- I en søgning med flere ord indsætter søgemaskinen automatisk ***og*** mellem ordene. Derfor får man kun resultater, hvor begge ord optræder.
    - Dvs. ikke artikler hvor han blot omtales som ***Schlüter***

Eksempel 2:

Svangerskabsafbrydelse OR fosterfordrivelse

- Giver resultater hvor mindst ét af ordene fremgår.

(svangerskabsafbrydelse OR fosterfordrivelse) AND debatindlæg

- Resultater hvor mindst ét af ordene fremgår, men ikke uden at ordet ***debatindlæg*** også fremgår.

(svangerskabsafbrydelse OR fosterfordrivelse OR abort) AND familyId:landogfolk1945

- Finder debatindlæg om svangerskabsafbrydelse eller fosterfordrivelse i Land og Folk.
- Liste over Avis-id : http://www.statsbiblioteket.dk/nationalbibliotek/mediestream-filer/avistiteloversigt/at_download/file

(svangerskabsafbrydelse OR fosterfordrivelse OR abort) AND familyId:landogfolk1945 py:1933

- Kun resultater fra 1933.

(svangerskabsafbrydelse OR fosterfordrivelse OR abort) AND familyId:landogfolk1945 py:[1933 TO 1935]

- Resultater fra årene 1933 til 1935.

(svangerskabsafbrydelse OR fosterfordrivelse) AND debatindlæg pu:ringkjøbing

- Giver resultaterne i aviser, der har Ringkjøbing som udgivelsessted.

Eksempel 3:

Klaus Riskær Petersen gav den rigtig gas på Venstres landsmøde 21.-23. september 1990 (måske var det kun to dage): Han sammenlignede det danske velfærdssamfund og Socialdemokratiet med alt det, som Østeuropa havde gjort oprør mod. Kan du finde det til mig?

(riskjær OR riskær) AND socialdemokrati* iso_date:[1990-09-20 TO 1990-09-25]

- Ved man derimod præcis hvilken dag man skal bruge så bruge man date-koden:

(riskjær OR riskær) AND socialdemokrati* date:(24. september 1990)

Eksempel 4:

Brug af kb.labs *Word2vec*:

Word2vec er en såkaldt un-supervised læringsalgoritme. Word2vec bygger på at ord som optræder i lignende sammenhænge, vil ligge tæt på hinanden i vector-rum. Dette kan blandt andet bruges til at få kvalificerede gæt på, hvordan et ord er blevet fejllæst af OCR. Korpusset er kun aviserne i årene 1900-2016, men fejllæsningerne som Word2vec spytter ud ser ud til også at kunne optimere en søgning i ældre fraktur-aviser.
http://labs.statsbiblioteket.dk/dsc/

Nearest words: vandværk Go

Analogy: man is to woman

#	Word
1	fjernvarmeforbrugere Search in Mediestream
2	trykforøgerstation Search in Mediestream
3	pumpestationen Search in Mediestream
4	vandurene Search in Mediestream
5	kraftvarme Search in Mediestream
6	fjernvarme Search in Mediestream
7	hovedvandledning Search in Mediestream
8	gadelys Search in Mediestream
9	elværk Search in Mediestream
10	afløb Search in Mediestream
11	renseanlæggets Search in Mediestream
12	vandvæl Search in Mediestream
13	kloakanlæg Search in Mediestream
14	vandvæi Search in Mediestream
15	trixtank Search in Mediestream

Her er vi interesserede i 12 og 14: ***vandvæl*****,** og ***vandvæi*****.** Disse to ord ligner fejllæsninger af ***vandværk*****.** Ved at samle lignede ord op fra liste kan man køre lignende søgning:

(vandværk* OR vandvæl* OR vandvæi* OR vandverk* OR vandvark*) py:[1878 TO 1900] familyId:aarhusstiftstidende

- Denne søgning, der kun er i frakturskrift (stifttidende bruger fraktur i perioden 1878 – 1900) giver 1014 resultater. Disse resultater er i høj grad fejllæsninger af *vandværk*. Til sammenligning får vi kun 384 hits ud af søgningen:

vandværk py:[1878 TO 1900] familyId:aarhusstiftstidende

Word2vec giver os altså kvalificerede bud på OCR-fejllæsninger.

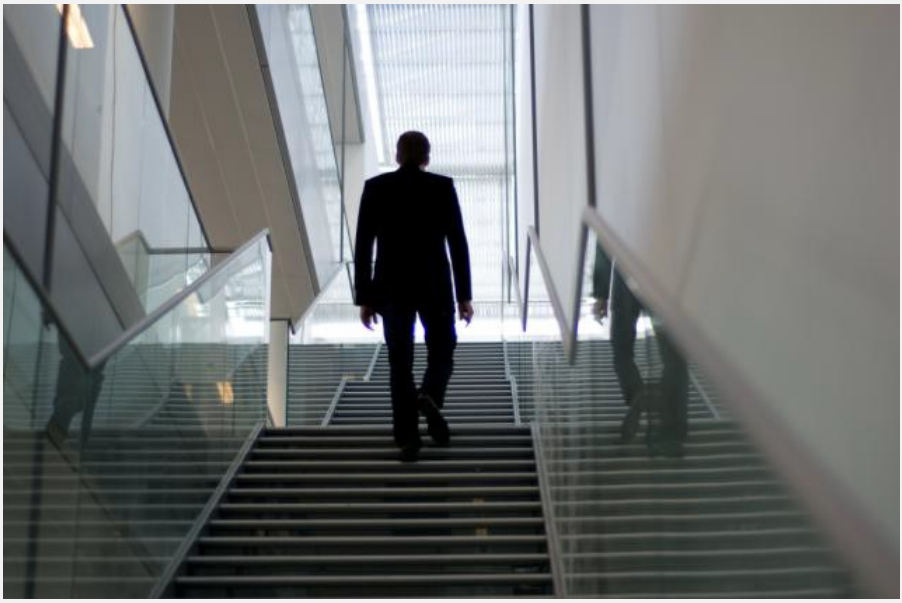
Udarbejdet af Max Odsbjerg Pedersen
*Aarhus Universitet**

*Basen *Mediestream* er udarbejdet af Statsbiblioteket i Aarhus, nu Det Kgl. Bibliotek, Aarhus. http://www2.statsbiblioteket.dk/mediestream/

Noter

Noter

KØBENHAVNS UNIVERSITETSBIBLIOTEK



Københavns Universitetsbibliotek, Søndre Campus

Karen Blixens Plads 7
2300 København S

kub.kb.dk/sc