# Hemuppgift 4

R för samhällsvetare. Stockholm universitet. Sommar 2020.

**Besvaras senast måndag 13 juli 17:00.**

1. Inlämningsuppgifter besvaras genom att ladda upp ett R-skript på Athena . Om du föredrar går det även bra att redovisa dina svar i ett worddokument, pdf, powerpoint eller dylikt. Döp filen till ”*Hemuppgift\_4.R*”. Du behöver inte ange ditt namn.
2. Det väsentliga är inte att dina resultat är snyggt presenterade. Huvudsaken är att du finner någon form av svar på frågan. Oavsett filformat ska det tydligt framgå vilka kommandon du använder för att lösa respektive uppgift. Skriv kommentarer i din kod (#).
3. Om du kör fast på någon fråga – lämna inte blankt. Gör så gott du kan och förklara kort vad som inte fungerar.

## Uppgift

1. Importera datatabellen ”regfor\_all.rds” till R.

Filen innehåller Sveriges regeringsförklaringar för varje femte år mellan 1980 och 2015. För att underlätta analysen har jag på förhand tagit bort alla siffror ur materialet.

1. Beräkna hur lång (antal ord) varje regeringsförklaring är. Visualisera resultatet i ett diagram. *Tips: börja med att tokensiera texten.*
2. Gör en TF-IDF-analys för att se vilka ord som var viktigast för *respektive år*. Visualisera resultatet i ett diagram. *Tips: detta kräver att du först skapar en kolumn för hur många gånger varje ord förekommer per år, sedan kan du gå vidare och göra tf-idf-analysen.*
3. Ett antal stoppord (”ska” och ”skall”) kommer att få stor vikt i TF-IDF-analysen. Ta bort dem och gör sedan om fråga tre. Det är okej att ta bort ännu fler stoppord om du vill, men ta som minst bort de två ovannämnda.
4. Importera ett sentimentlexikon och tillskriv varje ord sentimentpoäng enligt detta. Ge alla NA-observationer värdet 0. Gör sedan ett diagram över hur användandet av positiva och negativa ord skiljer sig åt mellan *åren 1980-2015*. *Tips: summera för varje år ordens poäng*.
5. Vilket *parti* var mest negativt respektive positivt sett till ordval om vi *inte* tar hänsyn till årtal? Ta hänsyn till att vissa partier använde sig av färre ord. Det innebär att sentimentspoängen måste divideras med antal ord per parti. *Tips: tre partier förekommer. Centerpartiet (Fälldin), Socialdemokraterna (Palme, Carlsson, Persson och Löfven) och Moderaterna (Reinfeldt)*.