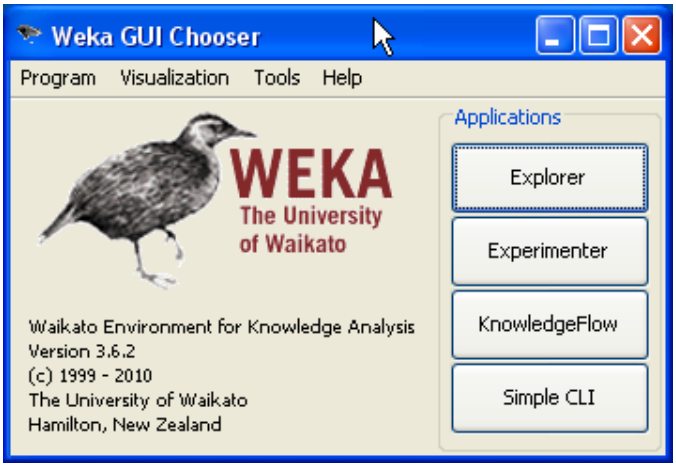
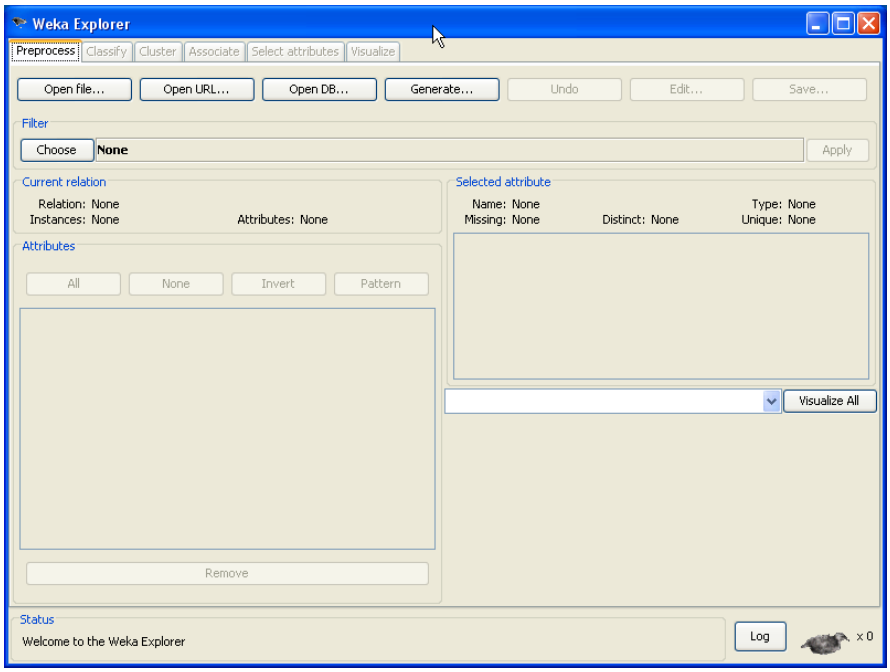
Att använda Weka för språkteknologiska problem

Systemet WEKA (Waikato Environment for Knowledge Acquisition) är en verktygslåda med olika maskininlärningsalgoritmer, metoder för att behandla indata, möjligheter för visualisering och utvärdering, mm. System är open-source och utvecklas huvudsakligen universitetet i Waikato, Nya Zeeland. Om du vill ladda ner systemet till egen dator kan hämta det här: http://www.cs.waikato.ac.nz/~ml/weka/index_downloading.html. Systemet finns också installerat på PC-datorerna i labbsalarna Vit och Magenta. Där startas Weka genom att **klicka på Start-knappen** längst ned till vänster på skärmen, välja **"All programs"** → **Weka 3.6.2** → **Weka 3.6"**.

När Weka startas öppnas ett fönster som ser ut så här:



Klicka på "Explorer". Då får ni upp ett nytt fönster som ser ut så här:



Här kan man nu öppna en fil med träningsdata genom att klicka på “Open file…”. Prova till exempel att ladda ner filen **test1.arff** från kurshemsidan och öppna den i WEKA. Öppna också test1.arff i ett vanligt textredigeringsprogram som Notepad eller Wordpad.

Om ni tittar på test1.arff i Notepad bör ni se detta:

```
@relation sentiment

@attribute phrase string
@attribute value {dummy, pos, neg, neut}

@data
“olyckan var framme”, neg
“bilen blev bara mos”, neg
“men mirakulöst nog klarade sig alla oskadda”, pos
“händelsen kommer nu att utredas”, neut
```

I detta lilla exempel försöker man beskriva sambandet mellan fraser och deras känsloladdning. Detta kodas som relationen “sentiment” som har två attribut, “phrase” som är en godtycklig sträng, och attributet “value” som kan anta ett av fyra värden: **dummy, pos, neg** och **neut** (värdet **dummy** är tillagt pga en bugg i Weka som gör att första värdet ignoreras i vissa situationer).

Under rubriken “@data” följer sedan exempel på positiva, negativa, och neutrala meningar.

I WEKA-gränssnittet kan ni klicka på de olika attributen under rubriken **Attributes**. Om man klickar på “value” får man se hur många meningar som är positiva eller negativa.

Man kan nu applicera olika filter på indata som förändrar representationen på olika sätt. Ett filter som är användbart för språkteknologiska applikationer är “StringToWordVector” (**Choose → filters → unsupervised → StringToWordVector**) , vilket omvandlar varje sträng till en bag-of-words. Ett attribut skapas för varje ord som förekommer bland alla exempelmeningarna, och värdet av detta attribut är en siffra som anger hur många gånger ordet förekommer i meningen. Välj detta filter, tryck **Apply** och spara under ett nytt namn med **Save…** Filen borde se ut så här:

```
@attribute sentiment {dummy,neg,pos,neut}
@attribute bara numeric
@attribute bilen numeric
@attribute blev numeric
@attribute framme numeric
@attribute mos numeric
@attribute olyckan numeric
@attribute var numeric
@attribute alla numeric
@attribute klarade numeric
@attribute men numeric
```

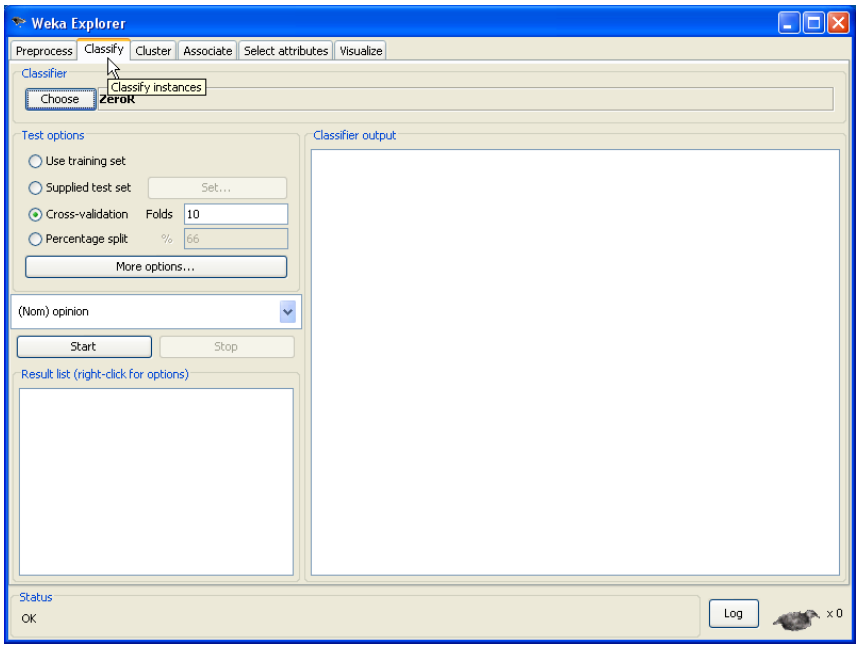
```
@attribute mirakulöst numeric
@attribute nog numeric
@attribute oskadda numeric
@attribute sig numeric
@attribute att numeric
@attribute händelsen numeric
@attribute kommer numeric
@attribute nu numeric
@attribute utredas numeric

@data
{0 neg,4 1,6 1,7 1}
{0 neg,1 1,2 1,3 1,5 1}
{0 pos,8 1,9 1,10 1,11 1,12 1,13 1,14 1}
{0 neut,15 1,16 1,17 1,18 1,19 1}
```

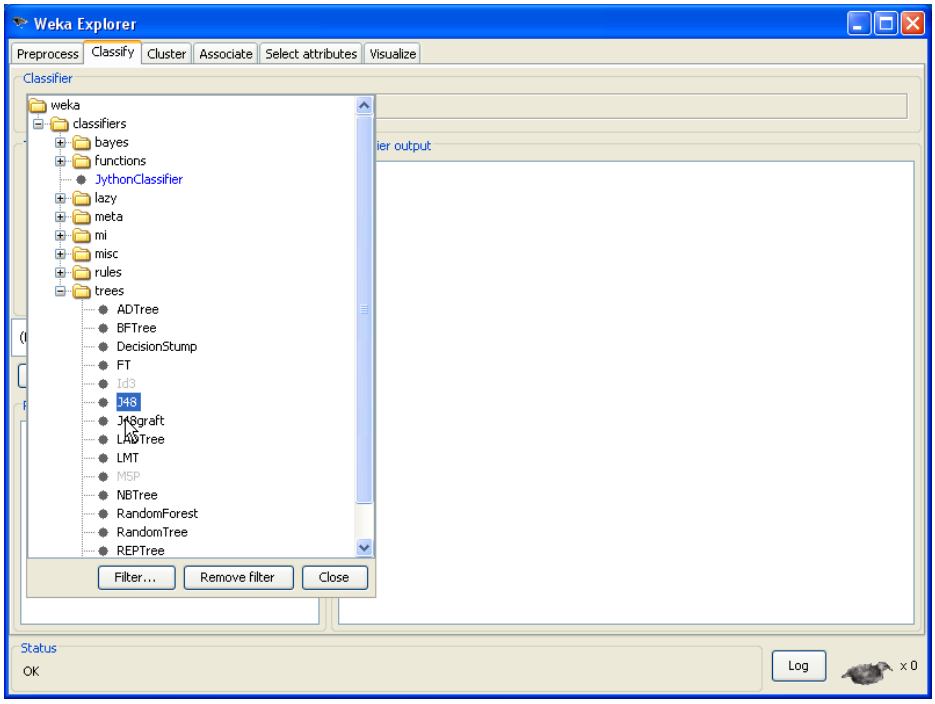
För att spara plats visas enbart attribut som inte är noll. Första exemplet bör t.ex. utläsas så här: Värdet av 0te attributet är **neg**, värdet av fjärde attributet (**framme**) är 1, värdet av sjätte attributet (**olyckan**)är 1, och värdet av sjunde attributet (**var**) är 1.

Man kan sedan transformera data en gång till med filtret NumericToBinary, vilket ger numeriska attribut värdet 1 ifall de har ett värde som är skilt från noll, och noll annars. Detta förvandlar alltså indata från en "bag-of-words" till en "set-of-words". Detta filter hittar du under **Choose → filters → unsupervised → NumericToBinary.**

Hittills har vi bara tittat på fliken **Preprocess** i WEKA-gränssnittet. Klicka nu på fliken **Classify**. Då borde du se något i stil med följande:

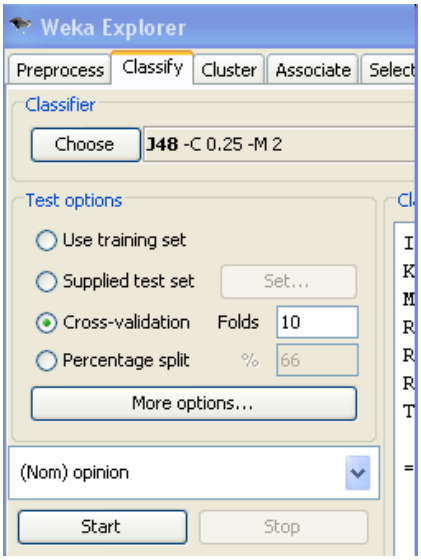


Genom att klicka på **Choose**-knappen kan man välja maskininlärningsmetod. Dokumentationen http://kent.dl.sourceforge.net/project/weka/documentation/3.6.x/WekaManual-3-6-2.pdf beskriver alla tillgängliga metoder.



Efter att man låtit WEKA bygga en klassificerare brukar man utvärdera resultaten. Detta går till så att WEKA kör ett antal testfall genom klassificeraren och jämför resultaten med facit. I vårt fall skulle man alltså som testfall använda ett antal fraser, se om klassificeraren föreslår **value=pos, value=neg** or **value=neut**, och jämföra med facit. WEKA kommer då att räkna ut hur stor andel av testfallen som blev korrekt klassificerade.

Man kan också välja hur resultaten ska utvärderas under **Test options**.



Det finns fyra möjligheter:

- **Use training set**. Testa med samma exempel som klassificeraren är tränad på. Detta anses som en dålig metod (risk för överträning!) men kan vara användbar när man bara har lite data.
- **Supplied test set**. Här kan man ange filnamn för en fil med testexempel.
- **Cross-validation**. Här körs ett antal pass med träning och test. Om folds=10 så används 90% av data till träning och 10% till test. Vid varje pass delas datamängden upp på ett nytt sätt, så att varje exempel används för träning 9 gånger och för test 1 gång.
- **Percentage split**. Här delas datamängden upp så att x% används för träning och 100-x% för test.

Välj **Use training set** och tryck på **Start**. WEKA kommer nu att skapa ett beslutsträd utifrån data och utvärdera trädet på träningsmängden. Utdata skrivs i fältet **Classifier output** till höger. Titta framför allt på två saker:

- Rubriken **Correctly Classified Instances** som talar om hur bra det gick.
- **Confusion Matrix** som talar om hur exempel blev klassificerade och hur de borde blivit klassificerade