ML Testing
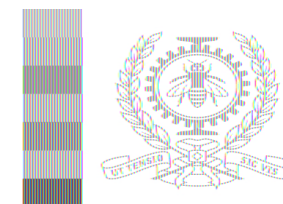
Laboratoire 4

Département de génie informatique et de génie logiciel
École Polytechnique de Montréal

Task 1 - start over again

- Let's get a more complex model a model using bigrams aka pairs of words thus

  - w1 w2 w3 w4 w5 generates

  - w1w2 w2w3 w3w4 w4w5 ….

  - hint: **CountVectorizer(*ngram_range*=(*1, 2*), *min_df*=*min_freq*)**

- Repeat all the tasks 1-11 from lab 3 the goal is indeed to decide if it worths using bigrams or unigrams are just fine

- for example compare confusion matrices, accuracy and AUC with dictionary size

- how about time ?

Task 2 - metamorphic properties

- Let's assume we use using bigrams aka pairs of words

- What metamorphic properties are still valid?

- Can we still randomly reorder words?

- Can we still shift all counter by 1?