To setup wikiprep-esa:

1. Python 2.7 is good. Python 3.X not good
   1. Python on windows needs VC++ for Python 2.7
   2. We need Python 64bit for large process!
   3. We need mySQLdb for python 2.7 64bit.
      1. <http://www.codegood.com/downloads> is good
2. scanLinks.py
   1. **Code is not very optimized! It is much better to output a file with all the rows and use LOAD DATA IN FILE. (much faster than a INSERT**
   2. Need to make work with split file+gzip:

count = 0

fname = args[0] + '.' + str(count).zfill(4) + ".gz"

while os.path.exists(fname):

print "working with file ", fname

f = gzip.open(fname,'r')

for doc in xmlwikiprep.read(f):

recordArticle(doc)

f.close()

count+=1

fname = args[0] + '.' + str(count).zfill(4) + ".gz"

1. scanData.py
   1. Need to customize DB connection details inside
   2. **My code Does addAnchors at the same time!** No need to run that
      1. I do it in-memory to haste things up. Process will need around 4-5Gb
   3. 2:50