PROJECT REPORT

NAME: Gopal Vishwakarma (dde911, @01688966)

PROBLEM STATEMENT:

To find the longest word chain (n-gram) among all the possible word chains in the dictionary.

PLATFORM USED:

The project was implemented using JAVA language.

DATA STRUCTURE USED:

This project used multiple data structures and methods implemented in JAVA.

1. Array List
2. HashMap
3. HashSet

ALGORITHM:

- Iteration method is used in this project.
- Step1: The dictionary.txt is being read by using buffered reader as a first step. Words are added in array list named as '*allwords'*.
- Step 2: All words are sorted.
- Step 3: Another data structure 'SetOfList' is created. In this, Different array list are created based on number of alphabets in words. i.e. for example, "aa" two letter word is added into 1st setoflist and "abc" three letter word will get added in 2nd setoflist and so on.
- Step 4:
    - The outermost *for* loop is used to iterate over all sets. (We got around 25 sets)
    - 2nd *for* loop is being used for to acquire words and initialize it as 'Key'.
    - 3rd *for* loop iterates over 'next adjacent set' which contains one extra letter.
    - 4th *for* loop gets all words one by one from next set (adjacent set which contains one extra letter words compared to previous set.
- Step 5: "*ContainsAllChars"* method is written to check all alphabets contained in words to be compared. "*addIntoMapOfChain"* method is used to create new chain of words and add words into it. Repeat Step 4 and Stpe 5 to get all chains.
- Simultaneously, these chains are added into 'n-gram-output.txt'. This process is carried on until all the end of the dictionary file.

COMPILE AND RUNNING PROGRAM:

1. Compiling program:

```
C:\Users\Gopal Vishwakarma\workspace\oxygen\DS_project\src>javac ngram.java

C:\Users\Gopal Vishwakarma\workspace\oxygen\DS_project\src>
```

2. Running Program:

```
C:\Users\Gopal Vishwakarma\workspace\oxygen\DS_project\src>java ngram
hello: Fri Dec 08 17:05:21 CST 2017
173528
Chain 1. aa -> aal -> aals -> ables -> abeles -> abelias -> abseiled -> abolished
Chain 2. ab -> aba -> abas -> abase -> abased -> abashed -> ambushed -> dumbheads -> husbandmen
Chain 3. ad -> add -> abed -> abide -> abided -> abiders -> abridges -> abridgers -> brigadiers -> abridgments -> abridgements
Chain 4. ae -> ace -> aced -> ached -> arched -> chaired -> archived -> chivareed
Chain 5. ag -> aga -> agar -> aargh -> aarrgh -> agoroth -> goatherd -> goatherds -> godfathers
Chain 6. ah -> aha -> ache -> ached -> arched -> chaired -> archived -> chivareed
Chain 7. ai -> aid -> acid -> acids -> alcids -> cladist -> citadels -> deistical -> acidulates -> colatitudes -> discountable
Chain 8. al -> aal -> aals -> ables -> abeles -> abelias -> abseiled -> abolished
Chain 9. am -> aim -> aims -> agism -> ageism -> ageisms -> armigers -> almsgiver -> almsgivers
```

SAMPLE RUNS AND OUTPUTS:

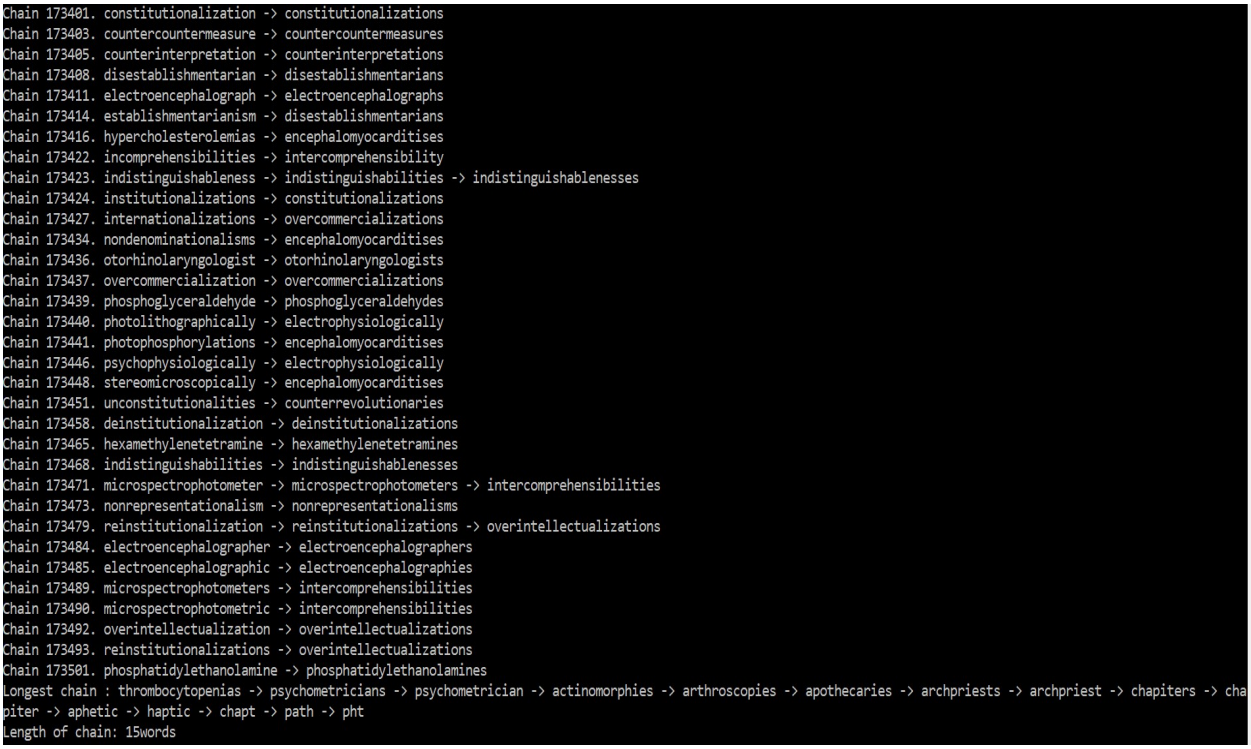
Following are the screen shots of the output obtained after running program for about 19 mins.

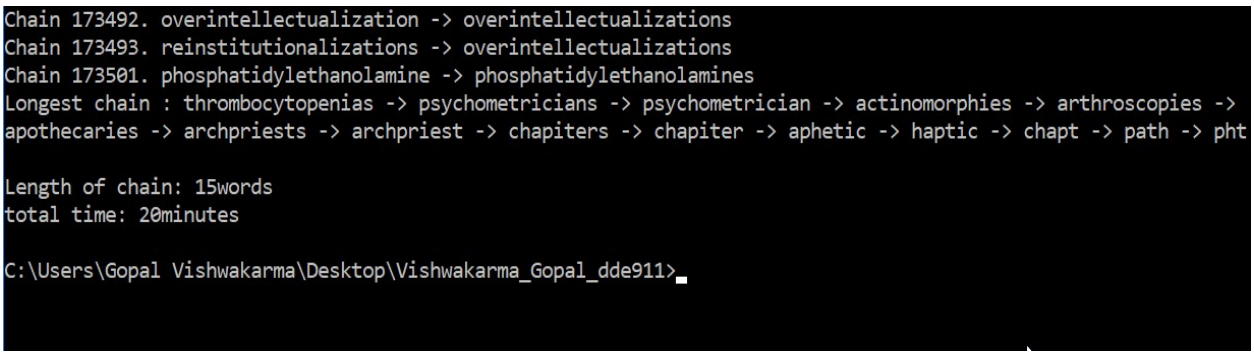
```
C:\Users\Gopal Vishwakarma\workspace\oxygen\DS_project\src>java ngram
hello: Fri Dec 08 18:45:36 CST 2017
173528
Chain 1. aa -> aal -> aals -> ables -> abeles -> abelias -> abseiled -> abolished
Chain 2. ab -> aba -> abas -> abase -> abased -> abashed -> ambushed -> dumbheads -> husbandmen
Chain 3. ad -> add -> abed -> abide -> abided -> abiders -> abridges -> abridgers -> brigadiers -> abridgments -> abridgements
Chain 4. ae -> ace -> aced -> ached -> arched -> chaired -> archived -> chivareed
Chain 5. ag -> aga -> agar -> aargh -> aarrgh -> agoroth -> goatherd -> goatherds -> godfathers
Chain 6. ah -> aha -> ache -> ached -> arched -> chaired -> archived -> chivareed
Chain 7. ai -> aid -> acid -> acids -> alcids -> cladist -> citadels -> deistical -> acidulates -> colatitudes -> discountable
Chain 8. al -> aal -> aals -> ables -> abeles -> abelias -> abseiled -> abolished
Chain 9. am -> aim -> aims -> agism -> ageism -> ageisms -> armigers -> almsgiver -> almsgivers
Chain 10. an -> ain -> agin -> acing -> aching -> arching -> archings -> scarphing -> purchasing
Chain 11. ar -> air -> abri -> abris -> airbus -> bariums -> biramous
Chain 12. as -> aas -> aals -> ables -> abeles -> abelias -> abseiled -> abolished
Chain 13. at -> act -> acta -> aceta -> abject
Chain 14. aw -> awa -> alow -> aglow -> logway -> dayglow -> dayglows
Chain 15. ax -> axe -> apex -> expat -> expats
Chain 16. ay -> aby -> ably -> badly -> baldly -> audibly
Chain 17. ba -> aba -> abas -> abase -> abased -> abashed -> ambushed -> dumbheads -> husbandmen
Chain 18. be -> bed -> abed -> abide -> abided -> abiders -> abridges -> abridgers -> brigadiers -> abridgments -> abridgements
Chain 19. bi -> bib -> abri -> abris -> airbus -> bariums -> biramous
Chain 20. bo -> abo -> abos -> ambos -> abmhos
Chain 21. by -> aby -> ably -> badly -> baldly -> audibly
Chain 22. de -> bed -> abed -> abide -> abided -> abiders -> abridges -> abridgers -> brigadiers -> abridgments -> abridgements
Chain 23. do -> ado -> ados -> adios -> adonis -> adjoins -> adjoints
Chain 24. ed -> bed -> abed -> abide -> abided -> abiders -> abridges -> abridgers -> brigadiers -> abridgments -> abridgements
Chain 25. ef -> eff -> alef -> alefs -> fables -> baffles -> bafflers -> balefires -> fiberglass
Chain 26. eh -> edh -> edhs -> ashed -> bashed -> abashed -> ambushed -> dumbheads -> husbandmen
Chain 27. el -> ale -> able -> abele -> abeles -> abelias -> abseiled -> abolished
Chain 28. em -> elm -> alme -> almeh -> almehs -> hamlets -> thermals
Chain 29. en -> ane -> acne -> acned -> acnode -> acnodes -> celadons -> canoodles -> adolescent -> adolescents -> adolescently -> considerately
Chain 30. er -> are -> acre -> acerb -> backer -> backers -> backrest -> backrests -> backbiters -> blacklister -> blacklisters
Chain 31. es -> eds -> beds -> based -> abased -> abashed -> ambushed -> dumbheads -> husbandmen
Chain 32. et -> ate -> abet -> abate -> abated -> abetted -> abdicate -> abdicated -> abstricted -> breadsticks
Chain 33. ex -> axe -> apex -> expat -> expats
Chain 34. fa -> aff -> afar -> afars -> afresh -> chafers -> chaffers -> chauffers -> chauffeurs -> ultrafiches
Chain 35. go -> ago -> agio -> agios -> amigos -> gliomas -> algorism -> algorisms -> algorithms
Chain 36. ha -> aha -> ache -> ached -> arched -> chaired -> archived -> chivareed
Chain 37. he -> edh -> edhs -> ashed -> bashed -> abashed -> ambushed -> dumbheads -> husbandmen
```

Chain 670. noh -> chon -> chino -> chinos -> chitons -> chitosan -> antishock -> mackintosh
Chain 671. nom -> mano -> amino -> ammino -> alimony -> palimony -> amylopsin -> amylopsins -> polynomials -> amylopectins
Chain 672. noo -> aeon -> aeons -> agones -> agonies -> agonised -> alongside -> desolating -> delegations -> desolatingly
Chain 673. nor -> born -> baron -> barong -> barongs -> begroans -> baronages -> abnegators -> baronetages -> beardtongues -> battlegrounds
Chain 674. nos -> cons -> canso -> acorns -> anchors -> chantors -> anchorets -> anchorites -> achondrites
Chain 675. not -> font -> fonts -> founts
Chain 676. now -> down -> adown -> onward -> onwards -> sandworm -> markdowns
Chain 677. nth -> hant -> chant -> canthi -> acanthi -> anorthic -> anchorite -> achondrite -> achondrites
Chain 678. nub -> bund -> bound -> abound -> abounds -> baudrons
Chain 679. nun -> anus -> ankus -> ankush -> kahunas -> ankushes -> lunkheads -> unshackled
Chain 680. nus -> anus -> ankus -> ankush -> kahunas -> ankushes -> lunkheads -> unshackled
Chain 681. nut -> aunt -> aunts -> cantus -> canthus -> acanthus -> ceanothus -> headcounts
Chain 682. oaf -> fado -> fados
Chain 683. oak -> amok -> amoks -> oakums
Chain 684. oar -> aero -> adore -> adored -> aborted -> broadest -> adsorbate -> abductores
Chain 685. oat -> alto -> allot -> abvolt -> abvolts
Chain 686. obe -> bode -> abode -> aboded -> abdomen -> abdomens
Chain 687. obi -> bios -> bison -> basion -> albinos -> coalbins -> balconies -> cobaltines -> bisectional -> celebrations -> elucubrations -> binocularities -> antitubercu
lous -> antituberculosis
Chain 688. oca -> arco -> acorn -> acorns -> anchors -> chantors -> anchorets -> anchorites -> achondrites
Chain 689. odd -> ados -> adios -> adonis -> adjoins -> adjoints
Chain 690. ode -> bode -> abode -> aboded -> abdomen -> abdomens
Chain 691. ods -> ados -> adios -> adonis -> adjoins -> adjoints
Chain 692. oes -> does -> bodes -> abodes -> albedos -> absolved
Chain 693. off -> boff -> befog -> befogs
Chain 694. oft -> coft -> croft -> crofts -> factors -> forecast -> factories -> factorizes
Chain 695. ohm -> holm -> holms -> holism -> holisms -> demolish -> halidomes
Chain 696. oho -> ahoy -> hoagy
Chain 697. ohs -> bosh -> hobos -> abhors -> barhops
Chain 698. oil -> boil -> aboil -> albino -> albinos -> coalbins -> balconies -> cobaltines -> bisectional -> celebrations -> elucubrations -> binocularities -> antitubercu
lous -> antituberculosis
Chain 699. oka -> amok -> amoks -> oakums
Chain 700. oke -> coke -> choke -> choked -> chocked -> cokehead -> blockhead -> blockheads
Chain 701. old -> bold -> blond -> blonde -> blonder -> banderol -> banderole -> banderoles -> borderlands -> adorableness -> banderilleros -> bildungsromane
Chain 702. ole -> aloe -> aloes -> aldose -> albedos -> absolved
Chain 703. oms -> doms -> demos -> demobs -> bosomed -> abdomens
Chain 704. one -> aeon -> aeons -> agones -> agonies -> agonised -> alongside -> desolating -> delegations -> desolatingly
Chain 705. ons -> cons -> canso -> acorns -> anchors -> chantors -> anchorets -> anchorites -> achondrites
Chain 706. ooh -> ahoy -> hoagy

Chain 161727. contiguousness -> centrifugations -> reconfigurations
Chain 161728. continuousness -> aluminosilicate -> aluminosilicates -> communicabilities -> commensurabilities
Chain 161729. contortionists -> abstractionisms -> collaborationism -> circumambulations -> commensurabilities
Chain 161730. contrabandists -> bronchodilators -> submitochondrial
Chain 161731. contrabassists -> abstractionisms -> collaborationism -> circumambulations -> commensurabilities
Chain 161732. contrabassoons -> abstractionisms -> collaborationism -> circumambulations -> commensurabilities
Chain 161733. contraceptions -> chronotherapies -> anthropocentrism -> anthropocentrisms
Chain 161734. contraceptives -> comparativeness -> overcompensating
Chain 161735. contractionary -> aerodynamicists -> thermodynamicist -> thermodynamicists -> aerothermodynamics
Chain 161736. contradictable -> bidirectionally -> decarboxylations
Chain 161737. contradictions -> achondroplastic
Chain 161738. contradictious -> cotransductions -> antireductionism -> antireductionisms -> adenocarcinomatous
Chain 161739. contraindicate -> adrenalectomies
Chain 161740. contraposition -> achondroplastic
Chain 161741. contrapositive -> comparativeness -> overcompensating
Chain 161742. contrapuntally -> hyperfunctional
Chain 161743. contrapuntists -> anticorruptions -> computerizations
Chain 161744. contrarinesses -> adrenalectomies
Chain 161745. contraventions -> coinvestigators -> magnetostrictive
Chain 161747. contritenesses -> adrenalectomies
Chain 161748. controllership -> controllerships -> electrophoresing -> neuropsychologist -> neuropsychologists
Chain 161749. controvertible -> convertibleness -> convertibilities -> convertiblenesses -> inconvertibilities
Chain 161750. contumaciously -> communistically -> commensurability
Chain 161751. contumeliously -> ethnomusicology
Chain 161752. convalescences -> anticonvulsives -> overcultivations -> revascularization -> revascularizations
Chain 161753. conventionally -> communicatively
Chain 161754. conventioneers -> coinvestigators -> magnetostrictive
Chain 161755. conversational -> controversially -> conversationally -> uncontroversially
Chain 161756. conversaziones -> overcentralizes
Chain 161757. convertaplanes -> overspeculating
Chain 161759. convertiplanes -> overspeculating
Chain 161760. convincingness -> coinvestigators -> magnetostrictive
Chain 161761. convivialities -> anticonvulsives -> overcultivations -> revascularization -> revascularizations
Chain 161762. convulsiveness -> anticonvulsives -> overcultivations -> revascularization -> revascularizations
Chain 161763. cooperationist -> chronotherapies -> anthropocentrism -> anthropocentrisms
Chain 161764. coordinateness -> adrenalectomies
Chain 161765. copartnerships -> chronotherapies -> anthropocentrism -> anthropocentrisms
Chain 161767. coprosperities -> archiepiscopate -> anthropocentrism -> anthropocentrisms
Chain 161768. coquettishness -> microtechniques
Chain 161769. coreligionists -> commiseratingly

Result screenshot is as below:





The screen shots shown are not only the outputs obtained, for the complete lists refer to the file n-gram-output.txt in the zip file.

COMPLEXITY ANALYSIS AND CONCLUSION:

- The algorithm designed works well for smaller inputs. For example, here the given dictionary is first sorted and divided into number of sets based on number of alphabets in a word.
- However, the real time is saved when current word compares with words in adjacent set only. This avoids iterating throughout dictionary.
- Here dictionary contains 173528 words. this program runs for approximately 19 mins to get longest Ngram chain.
- So, the conclusion is the algorithm is only efficient for smaller inputs and the complexity of the algorithm is high for the larger inputs.