Arbres Lexicogràfics Eficients

Professorat d'Algorísmia (GRAU-A)
Departament de Ciències de la Computació
Universitat Politècnica de Catalunya

Q1 2025–2026

Resum

Aquest projecte se centra en la construcció i ús d'estructures lexicogràfiques, com els tries i tries compactes (per exemple, els Patricia tries), per dur a terme operacions bàsiques de consulta, suggerència i anàlisi sobre dades textuals diverses. L'objectiu principal és comprendre el funcionament intern d'aquestes estructures, avaluar-ne l'eficiència i aplicar-les a escenaris reals de processament de textos.

Els estudiants hauran d'implementar diverses variants de tries, desenvolupar algorismes per a la cerca de paraules o fragments. A més, s'inclou una fase d'experimentació on es mesuraran mètriques de rendiment, com ara el temps de consulta o la memòria utilitzada, per contrastar les implementacions amb les propietats teòriques. El projecte ofereix una visió pràctica i aplicada de les estructures de dades prefixades, així com una oportunitat per treballar amb conjunts de dades reals.

1 Normes

El projecte es realitzarà en grups de **4 persones** (excepcionalment, en grups de 3 si hi ha una autorització expressa del professorat). Per formalitzar els grups, cal inscriure's al fitxer compartit `GRAU-A Projecte Equips (Q1, 25-26)`. En aquest document trobareu una columna per introduir un identificador d'equip (1, 2, 3, ...) i, a continuació, els cognoms, nom i subgrup de cadascun dels integrants, seguint l'exemple proporcionat. Aquesta inscripció s'ha de completar **abans del 30 de setembre de 2025**. Els estudiants que no hagin formalitzat un grup dins d'aquest termini es consideraran com a no participants en el projecte i, per tant, es considerarà que han abandonat l'assignatura.

El lliurament del projecte es farà exclusivament en línia mitjançant el **Racó FIB**. La data límit d'entrega serà a les 23:59 hores del dia **24 d'octubre de 2025**. Totes les comunicacions públiques relatives al projecte es realitzaran a través del *Racó FIB* o del canal de Slack `#projecte`. Durant el procés de correcció, el professorat pot contactar-vos per demanar aclariments o resoldre dubtes sobre el vostre treball.

2 Especificació del projecte

Aquest projecte té com a objectiu l'estudi i aplicació d'estructures lexicogràfiques, com els tries i radix tries, per a la gestió eficient de conjunts de paraules i l'anàlisi estadística de textos reals. A través de la construcció d'aquestes estructures i el seu ús en diferents escenaris, l'estudiant podrà explorar el comportament, eficiència i aplicacions pràctiques dels arbres lexicogràfics. El projecte es divideix en tres parts:

- **Part 1: Construcció de tries.** Implementació de tries lexicogràfiques, com els tries clàssics i radix tries, amb suport per a inserció i consulta eficient de paraules o frases.
- **Part 2: Algorismes de cerca de paraules.** Donat un fitxer de text i una consulta (una paraula, un sufix, una frase o un fragment), cal identificar en quines línies del text apareix.
- **Part 3: Experimentació.** Anàlisi experimental del comportament de les estructures desenvolupades, tant en temps com en espai, aplicades a conjunts de dades reals. Comparació empírica amb les cotes asimptòtiques previstes.

Aquest document és intencionadament vague. Hi ha molta bibliografia accessible al respecte i no us costarà gens trobar-ne informació. No es tracta només de seguir unes instruccions, sinó de prendre decisions de disseny, justificar-les, i documentar-les adequadament. S'espera que l'alumnat investigui fonts externes, conegui diverses alternatives algorísmiques i les valori críticament. Això inclou l'elecció dels mètodes de cerca, el tractament dels textos, els criteris de similitud, la gestió de la memòria i el disseny experimental. També és part fonamental del projecte desenvolupar una metodologia clara, comprendre les limitacions dels algorismes implementats i reflexionar sobre els resultats obtinguts.

Part 1: Construcció de tries.

Un *trie* (també conegut com a arbre digital o arbre de prefixos) és una estructura de dades que permet emmagatzemar i cercar paraules o claus textuals de manera eficient. A diferència d'un arbre binari de cerca, els nodes d'un trie no contenen directament la clau associada, sinó que la seva posició dins de l'arbre determina aquesta clau. Les connexions entre nodes es defineixen a partir dels caràcters individuals de la clau, no pas de la clau completa.

Les seves aplicacions són àmplies. Els tries són especialment efectives en tasques com l'autocompletat, la correcció ortogràfica o el encaminament d'adreces IP, ja que ofereixen avantatges respecte a les taules de dispersió (hash tables) gràcies a la seva organització basada en prefixos i a l'absència de col·lisions de hash.

Suposem que disposem d'un alfabet Σ de n símbols possibles. Un trie es pot representar de diverses maneres, cadascuna amb compromisos diferents entre l'ús de memòria i l'eficiència de les operacions. La representació més senzilla d'un trie és la d'un arbre arrelat on cada vèrtex representa un prefix d'alguna paraula del conjunt, i cada aresta està etiquetada amb un símbol de l'alfabet Σ. Cada camí des de l'arrel fins a un vèrtex full correspon a una paraula completa del conjunt. La Figura 1 mostra un exemple senzill d'un trie.

Representar cada node amb un vector de punters pot consumir una gran quantitat d'espai, ja que la majoria d'entrades sovint són buides (Nulls). Per reduir l'ús de memòria, es pot utilitzar una llista enllaçada per representar només els enllaços no buits, tot i que això pot afectar negativament el temps d'accés. A més, tècniques com la reducció de l'alfabet poden ajudar a disminuir l'espai necessari, reinterpretant les cadenes originals com a cadenes més llargues sobre un alfabet més petit.

Un *radix tree*, també conegut com a *trie comprimit*, és una variant optimitzada en espai del trie clàssic, on qualsevol node amb un únic fill es fusiona amb el seu pare. L'eliminació de branques innecessàries en nodes amb un sol descendent permet millorar tant l'eficiència espacial com el temps d'accés. Els *patricia trees* són una implementació particular de tries comprimides binàries, que utilitzen la codificació binària de les claus de text. Cada node conté un índex (conegut com a *skip number*) que indica la posició de bifurcació, amb l'objectiu d'evitar subarbres buits durant el recorregut. Una implementació naïf d'un trie pot consumir una gran quantitat de memòria, degut a la dispersió de les claus i al nombre elevat de fulles. En aquests casos, els tries comprimides com el patrícia poden ser molt més eficients.

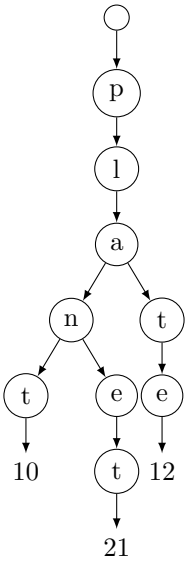

Figura 1: Exemple de trie per a {`plant`,10} , {`planet`,21} i {`plate`,12}.

En aquesta part es demana:

- Reviseu els textos acadèmics i la literatura sobre els tries i les seves implementacions eficients. Fer un estudi i un resum de les seves propietats, incloent-hi les cotes asimptòtiques tant de temps com d'espai, per a les diferents variants dels tries.
- Implementeu la representació estàndard d'un trie, així com les implementacions que considereu més interessants i eficients. Per a cada implementació seleccionada, cal desenvolupar tots els procediments necessaris per crear tries buits, inserir paraules, frases o sufixos, i preparar-los per a consultes eficients.

Per tal de facilitar la feina i ajudar-vos a començar, es proporcionen unes dades de referència que inclouen diversos textos de domini públic amb contingut escrit en diferents alfabets (per exemple, l'alfabet ASCII o seqüències d'ADN amb només quatre símbols). Aquest conjunt de dades constitueix una base recomanada per iniciar les proves, però no és limitatiu: els alumnes poden proposar altres textos que considerin rellevants per a la seva proposta.

Fitxer	**Descripció**	**Mida**	**Font**
`alice_wonderland.txt`	Llibre *Alice in Wonderland*	171,0 KB	Enllaç
`moby_dick.txt`	Llibre *Moby Dick*	1,3 MB	Enllaç
`words_alpha.txt`	Diccionari d'anglès	4,1 MB	Enllaç
`dna_genome.txt`	Genoma d'ADN (llevat)	12,0 MB	Enllaç
`wikipedia_titles.gz`	Títols d'articles de la Wikipeadia	390,0 MB	Enllaç

Taula 1: Fitxers de dades recomanats per començar el projecte (no exclusius).

Part 2: Algorismes de cerca de paraules o frases.

Un trie permet dur a terme diverses operacions fonamentals sobre conjunts de paraules, com la cerca exacta, l'ordenació lèxica, la detecció de prefixos compartits (com el *longest common prefix*) o la generació d'estructures de suggeriment basades en l'inici d'una paraula. Aquestes estructures esdevenen especialment útils en aplicacions de cerca, autocompletat o anàlisi estadística de textos. En aquest projecte, ens centrarem en l'ús de tries per a realitzar cerques eficients de paraules i frases dins d'un text donat.

Un dels aspectes clau d'aquesta part és definir clarament què es vol indexar: es pot optar per construir el trie únicament a partir de paraules senceres (separades per espais o signes de puntuació) o bé indexar tots els sufixos del text per tal de poder localitzar frases o fragments arbitràriament posicionats. Aquesta segona estratègia, més costosa en memòria, permet cerques més flexibles i pot capturar patrons més complexos, com fragments que travessen diverses paraules o apareixen en posicions no prefixades.

Per dur a terme una cerca dins del trie, cal recórrer l'estructura des de l'arrel seguint els caràcters de la consulta fins a trobar una coincidència completa. En cas que no existeixi cap branca que coincideixi amb la consulta, es pot interrompre anticipadament el recorregut i concloure que no hi ha cap resultat. Ara bé, una operació especialment rellevant en contextos d'ús real és l'autocompletat. En aquest cas, el trie es recorre parcialment segons el prefix de la consulta, i posteriorment es recuperen totes les paraules (o frases) que comencen per aquest prefix (tant de manera exacta com aproximada). Aquestes paraules es poden ordenar segons diversos criteris, com ara la freqüència, la longitud o la distància lèxica respecte de la consulta, amb l'objectiu d'oferir suggeriments rellevants i útils.

En aquesta part es demana:

- Implementeu eficientment la cerca d'una paraula o frase dins d'un trie construït a partir d'un fitxer de text. Per a cada consulta (una paraula, un sufix, una frase o un fragment), cal indicar en quines línies del text apareix.
- **(Extra)** En els casos en què la consulta no aparegui al text, implementeu un sistema d'autocompletat que, donada aquesta entrada parcial, retorni les cinc paraules o frases més properes, segons un criteri justificat.

Part 3: Experimentació

L'objectiu d'aquesta part del projecte és analitzar empíricament el rendiment de les estructures de tries desenvolupades a les parts anteriors. L'estudiant haurà de dissenyar un conjunt d'experiments per quantificar l'eficiència tant de la construcció com de les operacions de cerca, utilitzant conjunts de dades textuals diversos i contrastant els resultats amb les propietats teòriques conegudes.

L'anàlisi experimental haurà de tenir en compte com varien els resultats segons la mida del conjunt de dades, el tipus d'alfabet utilitzat (per exemple, ASCII o ADN), i la longitud de les consultes realitzades. Per això, es recomana realitzar proves amb consultes curtes (de pocs caràcters), mitjanes (com paraules senceres) i llargues (fragments o frases), i observar si la resposta de les diferents implementacions del trie es manté estable o canvia significativament.

Durant la construcció de les estructures, es pot estudiar el nombre total de nodes creats, la profunditat mitjana de l'arbre i l'espai de memòria que ocupa. Pel que fa a les cerques, serà rellevant mesurar el nombre de nodes visitats, el temps de resposta i el comportament del sistema quan s'utilitzen funcions com l'autocompletat. Aquestes dades permetran comparar el rendiment

de les variants implementades i determinar quines són més adequades segons el tipus d'entrada o d'aplicació.

Finalment, caldrà resumir els resultats de forma clara i estructurada. S'espera que es presenti els resultats amb gràfics, figures o taules que ajudin a visualitzar les diferències observades. També és important incloure una anàlisi crítica que expliqui si els resultats concorden amb les expectatives teòriques i quines conclusions es poden extreure del comportament observat.

3 Avaluació del projecte

Aquest document és intencionadament vague per tal d'estimular la vostra capacitat d'investigació i presa de decisions. S'espera que investigueu pel vostre compte totes les tècniques algorísmiques i models esmentats, i que escolliu les solucions que considereu més adequades per al vostre enfocament. Recordeu que cal implementar com a mínim dues versions d'un trie: una implementació bàsica i una altra optimitzada o més eficient. Un cop completats els experiments i recopilades totes les dades experimentals, heu de preparar un informe en què es descrigui de manera clara i sintètica la vostra implementació dels algorismes.

El nivell de sofisticació i l'esforç dedicat a la pràctica és opcional, però es tindrà en compte de manera positiva en l'avaluació del projecte. Es valorarà favorablement que implementeu més variants i algorismes, introduïu mesures internes addicionals i realitzeu els corresponents experiments i comparatives. Per exemple, no és suficient implementar una nova mètrica interna de qualitat si després no es fan proves experimentals ni es compara amb les altres mesures desenvolupades, o si no es descriu adequadament a l'apartat corresponent ni s'hi inclouen les referències bibliogràfiques pertinents. També es considerarà un punt a favor que dissenyeu nous experiments que permetin explicar el comportament observat dels algorismes en funció de les característiques dels conjunts de dades utilitzats.

És important estructurar l'informe diferenciant la part comuna (estructures bàsiques, operacions fonamentals, etc.) de les variants que hagueu desenvolupat, evitant repeticions innecessàries. L'ús de figures, esquemes i diagrames per il·lustrar el funcionament dels algorismes, l'estructura de les dades i els resultats obtinguts és altament recomanable. En canvi, incloure codi detallat en el cos principal de l'informe no és apropiat. Si voleu fer referència a parts específiques del vostre codi, podeu incloure pseudocodi o fragments concrets als apèndixs. L'informe principal no pot excedir les 10 pàgines, la qual cosa us obligarà a sintetitzar la informació, prioritzar els resultats més rellevants i evitar explicacions innecessàriament verboses. Recordeu que generar molt de text no implica una millor nota, i que no podeu confiar cegament en eines com ChatGPT per elaborar el vostre informe: el que es valora és la vostra capacitat d'anàlisi, síntesi i comunicació tècnica.

Caldrà utilitzar LaTeX per redactar l'informe del projecte. Per generar gràfiques, podeu emprar qualsevol dels nombrosos paquets disponibles per a LaTeX (especialment `tikz` i `pgf`) o bé fer servir programari extern com ara `gnuplot`, generant les imatges de manera independent i inserint-les posteriorment a l'informe. **Podeu trobar la plantilla LaTeX per començar a treballar en el següent** enllaç.

Tingueu en compte que la documentació lliurada ha de permetre avaluar el nivell d'assoliment de la competència transversal *capacitat d'autoaprenentatge*, que també forma part de l'avaluació del projecte. En el marc d'aquest projecte, hi ha múltiples aspectes rellevants relacionats amb aquesta competència: des de l'estudi de noves tècniques i models algorísmics, fins al disseny i execució d'experiments, així com la capacitat per documentar adequadament el treball realitzat amb criteris tècnics i científics.

La qualificació final del projecte reflectirà la qualitat de l'aprenentatge demostrat, l'expe-

rimentació realitzada i la documentació lliurada. El codi presentat haurà de ser funcional i correcte, però només tindrà un pes reduït en la nota global.

4 Detalls de l'entrega

Caldrà lliurar un fitxer comprimit (`.zip`), mitjançant el *Racó*, que contingui el codi font de tots els programes desenvolupats, i tots els fitxers addicionals que puguin ser necessaris per a la compilació i execució de cadascun d'ells, així com les instruccions per fer-los anar. Apart dels programes, s'ha d'incloure l'informe del projecte en PDF. Anomeneu aquest fitxer `InformeEquipID.pdf` (on `ID` és l'identificador del vostre equip de treball). No seguir aquestes indicacions a l'entrega podria tenir penalització a la qualificació.

No cal incloure els fitxers d'entrada utilitzats en els experiments, però convé incloure els fitxers amb totes les dades experimentals obtingudes.

Bibliografia recomanada

- Sedgewick, Robert, and Kevin Wayne. *Algorithms*, Addison-Wesley. [Section 5.3: Tries]. `https://algs4.cs.princeton.edu/lectures/keynote/52Tries.pdf`
- Mehta, Dinesh P., and Sartaj Sahni. *Handbook of Data Structures and Applications*, Chapman and Hall/CRC. [Chapter 28: Tries].
- Knuth, Donald E. *The Art of Computer Programming, Volume 3: Sorting and Searching*, Addison-Wesley. [Section 6.3: Digital Searching].